BMC Evolutionary Biology

Evolutionary rate depends on number of protein-protein interactions independently of gene expression level

BMC Evolutionary Biology 2004, 4:13

Hunter B Fraser (hunter@ocf.berkeley.edu)
Aaron E Hirsh (aehirsh@stanford.edu)

ISSN 1471-2148

Article type Debate

Submission date 11 Dec 2003

Acceptance date 27 May 2004

Publication date 27 May 2004

Article URL http://www.biomedcentral.com/1471-2148/4/13

Evolutionary rate depends on number of protein-protein interactions independently of gene expression level

Hunter B. Fraser[1]* and Aaron E. Hirsh[2]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA, 94720, USA. [2]Department of Biological Sciences, Stanford University, Stanford, CA, 94305, USA.

Email: Hunter B. Fraser – hunter@ocf.berkeley.edu; Aaron E. Hirsh - aehirsh@stanford.edu

*Corresponding author

Abstract

Background: Whether or not a protein's number of physical interactions with other proteins plays a role in determining its rate of evolution has been a contentious issue. A recent analysis suggested that the observed correlation between number of interactions and evolutionary rate may be due to experimental biases in high-throughput protein interaction data sets.

Discussion: The number of interactions per protein, as measured by some protein interaction data sets, shows no correlation with evolutionary rate. Other data sets, however, do reveal a relationship. Furthermore, even when experimental biases of these data sets are taken into account, a real correlation between number of interactions and evolutionary rate appears to exist.

Summary: A strong and significant correlation between a protein's number of interactions and evolutionary rate is apparent for interaction data from some studies. The extremely low agreement between different protein interaction data sets indicates that interaction data are still of low coverage and/or quality. These limitations may explain why some data sets reveal no correlation with evolutionary rates.

Background

Over twenty-five years ago, a number of authors suggested that a protein's rate of evolution should decrease with the number of molecular interactions in which it participates [1-3]. The rationale behind this prediction was that additional interactions impose functional constraints on otherwise relatively unconstrained residues, such as those on the surface of the protein. Thus, other things being equal, a protein with more interactions would evolve more slowly. This prediction was recently corroborated by us, in the form of a negative correlation between a protein's rate of evolution and the number of other proteins with which it interacts [4]. While other authors have questioned the existence of this relationship [5], we later showed that in their analysis, the absence of a correlation was due to the particular protein interaction data that they used; when all data sets available at that time were used, a very strong and statistically significant correlation was apparent [6].

In a recent, thorough analysis of protein interaction data sets, Bloom and Adami have questioned whether the correlation between number of protein interactions and evolutionary rate is independent of gene expression level [7]. While we agree that the results of Bloom and Adami show quite convincingly that an association between expression and number of interactions contributes significantly to the correlation between interactions and evolutionary rate, we believe that two of their conclusions are unwarranted. First, it is not yet clear that the association between expression and number of protein interactions is due exclusively to experimental biases rather than real properties of the organism. Second, current results do not indicate that the correlation between interactions and evolutionary rate is entirely due to the association between expression and evolutionary rate. In this work, we argue that their conclusions represent an over-extension of their analyses, and also provide further analyses demonstrating that a protein's number of interactions does indeed influence its rate of evolution, independently of its expression level.

Discussion

Critique of Bloom and Adami

Bloom and Adami [7] tested protein interaction data from seven methods (two experimental and five computational) individually for correlations between the number of protein interactions and protein evolutionary rates, while statistically controlling for gene expression levels. They found that only in the two interaction data sets generated using mass spectrometry was there a strongly significant correlation between the number of protein interactions and evolutionary rate independent of expression levels. In protein interaction data sets generated by the computational methods of gene co-occurrence and gene neighborhood, a weakly significant correlation between number of interactions and evolutionary rate remained when expression levels were statistically controlled [7]. Despite the inability of expression levels to account for the correlation between number of interactions and evolutionary rate in these data sets, Bloom and Adami argued that expression levels completely explain the correlation between number of interactions and evolutionary rate, and that they failed to see this in the partial correlations because the partial correlations did not completely control for expression levels. To explain why partial correlations were unable to completely control for expression levels, Bloom and Adami suggested that their expression data (measured by DNA microarrays and codon bias) are imprecise.

While we agree with Bloom and Adami that current codon usage and expression data do not measure expression levels with perfect precision, we do not believe that their interpretation is supported by the evidence. If one is to consider the quality of each of the types of data involved in calculation of the partial correlations—expression data, evolutionary rate data, and interaction data—there is no question that the least reliable of the three are the interaction data. This can be seen in many ways, the simplest of which is the nearly nonexistent overlap between different high-throughput protein interaction data sets [8]. Regardless of whether this small overlap is predominantly due to false positives, false negatives, or simply incomplete coverage, the fact is that the two independent expression data sets used by Bloom and Adami show much better agreement than any two high-throughput interaction data sets in existence. (As reported by Bloom and Adami [7], their two expression data sets are correlated with Spearman rank $r = 0.62$; in contrast, the correlation between number of interactions per protein in two of the most comprehensive and highest quality high-throughput interaction data sets [9, 10] is only 0.12, and correlations between most other protein interaction data sets are weaker or even negative [11]). Expression data, we may conclude, are of significantly higher quality and/or coverage than currently available interaction data. Therefore, if one is to invoke poor data as an explanation for not observing some particular outcome of the analysis, then it should be invoked to explain why the correlations involving protein interactions are not any stronger than they presently are. More generally, if the precedent set by Bloom and Adami were to be followed, then any variable A that only partially weakens a correlation between two other variables B and C when it is statistically controlled for could be claimed to be completely responsible for the correlation between B and C, if the values of A are not known with perfect precision. While it is certainly always possible that A will completely account for the correlation between B and C when it is known with more precision, this remains speculative in the absence of any supporting evidence.

As further evidence that the correlation between number of interactions and evolutionary rate is mediated by expression level, Bloom and Adami [7] showed that only in the interaction data sets in which the proteins with many interactions are highly expressed is there a significant negative correlation between number of interactions and evolutionary rate. Working under the assumption that the observed relationship between number of interactions and level of expression is an experimental artifact, Bloom and Adami suggested that the correlation between number of interactions and evolutionary rate is due to an experimental bias toward the detection of many interactions for highly expressed proteins. However, a simple alternative explanation must also be considered: it is entirely possible that highly expressed genes do tend to have more protein interactions than weakly expressed genes. Indeed, in addition to being found in yeast, a positive correlation between expression level and number of interactions has been reported in other organisms as well, using protein interaction detection methods (such as yeast 2-hybrid) which Bloom and Adami believe are unbiased with respect to expression levels [12]. If more highly expressed proteins do tend to participate in more protein interactions, one would expect to observe precisely the pattern of correlation coefficients Bloom and Adami report. Specifically, interaction datasets of sufficiently high coverage and accuracy would reveal the (real) relationship between expression and number of interactions, as well as the relationship between evolutionary rate and number of interactions. In contrast, less accurate or complete datasets would show no such relationships. As evidence against this idea, Bloom and Adami state that Jordan *et al.* [5] "observed no significant correlation between evolutionary rate and the number of interactions when they used a set of manually curated interactions that might be expected to be of higher accuracy than those from any single high–throughput method." While it is true that Jordan *et al*. did not observe a significant correlation, it is not true that they relied on a set of manually curated interactions. As we previously pointed out [6], approximately half of the interactions in the list used by Jordan *et al*. (after duplicate interactions were removed [13]) were from the high-throughput yeast 2-hybrid screen of Uetz *et al*. [14], which has been shown to be one of the least reliable high-throughput protein interaction data sets in existence [8].

Finally, Bloom and Adami criticized the biophysical explanation we proposed [4] to explain why proteins with many interactions would tend to evolve slowly. They stated that "there is no obvious reason why residues involved in intermolecular contacts should be more evolutionary [sic] constrained than other residues with the same number of intramolecular contacts" [7]. While this is true, it is not directly relevant to our original proposal, which was that "proteins with more interactions could evolve more slowly because a greater proportion of the protein is involved in protein functions" [4]. Our proposal was not that intermolecular contacts impose more stringent constraints than intramolecular contacts, but rather that additional interactions could impose constraints on sites that are otherwise relatively unconstrained, such as residues on the surface of a polypeptide. Thus the critique presented by Bloom and Adami has no bearing on the hypothesis we proposed.

Additional analysis of the data

A simple statistical method for examining the relationship between two variables (e.g., number of interactions, I and rate of evolution, E), while partially controlling for a third, potentially related variable (e.g., gene expression, A), is to divide the dataset into quantiles according to the controlled variable. This reduces the variance of the controlled variable relative to the other variables within each quantile, resulting in partial statistical control. This approach is complementary to partial correlation in that it the two methods can be combined, and division of the dataset into bins allows one to investigate the consistency or variation of relationships across quantiles. To emphasize that current data do not indicate that the relationship between evolutionary rate and number of interactions in mass spectrometry data is entirely mediated by expression levels, we present here a simple binning and partial correlation analysis of mass spectrometry [9], expression, and evolutionary rate data. It bears restating here that correlations among separate datasets indicate that interaction data are far less accurate than expression data; therefore, noise and other limitations of data should be expected to reduce the estimated strength of the relationship between number of interactions and evolutionary rate more than they reduce the strength of the relationship between expression levels and evolutionary rate.

As Bloom and Adami [7] noted, the proteins that are chosen to be tagged and overexpressed in mass spectrometry studies are subject to an ascertainment bias. For this reason, we used only the untagged data. We used the expression data of Wang *et al*. [15], which was produced from more replicates than other available expression datasets and, unlike the data used by Bloom and Adami [16], was not accidentally measured in an aneuploid strain of yeast [17]. For evolutionary rate data we used dN/dS values calculated from four species of the *Saccharomyces* genus, with a correction for the effect of codon bias on dS (Hirsh AE, Fraser HB, and Wall DP, submitted). We used Spearman's rank correlation for all analyses.

Before dividing the dataset into quantiles according to expression level, we measured the strength of the correlation between number of interactions and evolutionary rate for all 555 genes for which we had interaction, evolutionary rate, and expression data. The correlation between number of interactions (I) and evolutionary rate (E) was quite strong, even when controlling for expression ($r_{\mathrm{EI}} = -0.403$, $p = 5\mathrm{x}10^{-23}$; $r_{\mathrm{EI.A}} = -0.277$, $p = 3\mathrm{x}10^{-11}$; Table 1, row 1, column 1). We then partitioned the dataset into quantiles according to expression levels and calculated r_{EI} and

$r_{\mathrm{EI.A}}$ within each bin. We present results using two, three, four, and five bins. In every bin, the correlation between number of interactions and evolutionary rate is significant, even after controlling for expression levels. Perhaps even more importantly, controlling for expression levels actually *strengthens* the correlation between number of interactions and expression level in three of the bins (Table 1, underlined). In one of these bins, controlling for expression levels results in more than a two-fold improvement in the *p*-value of the correlation. In order for inaccurate expression data to explain this result, the expression data in those three bins would not only have to be noisy—they would have to be negatively correlated with the true expression levels of those genes. Since this is quite unlikely to be the case, we believe the most parsimonious explanation is that the number of interaction partners a protein has is correlated with its evolutionary rate independently of its expression level.

Summary

We agree with Bloom and Adami [7] that the quality of high-throughput protein interaction data sets is quite variable, and that some show a correlation with evolutionary rates while others do not. However we do not believe that expression levels can account for this correlation in all data sets. To support this position, we showed that limitations of the data are likely to weaken the apparent effect of number of interactions more than they weaken the apparent effect of expression. Therefore, Bloom and Adami's suggestion that the significant contribution of expression to the relationship between number of interactions and evolutionary rate should be interpreted to mean that expression is entirely responsible for this relationship seems unwarranted. To emphasize that the measurable effect of number of interactions on evolutionary rate remains highly significant even when controlling for expression, we presented a re-analysis of mass spectrometry interaction data. Across quantiles of expression, the relationship between number of interactions and evolutionary rate, controlling for expression levels, was significant. In several quantiles, controlling for expression actually strengthened the relationship between number of interactions and evolutionary rate.

Bloom and Adami's thorough analysis shows, above all, that large-scale data sets remain woefully noisy and incomplete. While it remains possible that expression levels will ultimately account for the correlation between number of interactions and evolutionary rate once more accurate expression data are published, we find it far more likely that the vast majority of improvement will be in protein interaction data. In any case, it will be interesting to see what relationships emerge as more (and higher quality) functional genomic data are produced.

Author contributions:

HBF performed the analyses. HBF and AEH wrote the manuscript.

References:

1. Dickerson RE: **The structures of cytochrome c and the rates of molecular evolution.** *J. Mol. Evol.* 1971, 1:26-45.
2. Ingram VM: **Gene evolution and the haemoglobins**. *Nature* 1961, 189:704-708.
3. Wilson AC, Carlson SS White TJ: **Biochemical evolution.** *Ann. Rev. Biochem.* 1977, 46:573-639.
4. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network**. *Science* 2002, 296:750-752.

5. Jordan IK, Wolf YI, Koonin EV: **No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly**. *BMC Evolutionary Biology* 2003, 3:1.

6. Fraser HB, Wall DP, Hirsh AE: **A simple dependence between protein evolution rate and the number of protein-protein interactions.** *BMC Evolutionary Biology* 2003, 3:11.

7. Bloom JD, Adami C: **Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets.** *BMC Evolutionary Biology* 2003, 3:21.

8. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, 417:399-403.

9. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al*.: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry**. *Nature* 2002, 415:180-183.

10. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al*.: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, 415:141-147.

11. Hoffman R,Valencia A: **Protein Interaction: same network, different hubs.** *Trends in Genetics* 2003, 19:681-683.

12. Cutter AD, Payseur BA, Salcedo T, Estes AM, Good JM, Wood E, Hartl T, Maughan H, Strempel J, Wang B, *et al*.: **Molecular Correlates of Genes Exhibiting RNAi Phenotypes in *Caenorhabditis elegans*.** *Genome Research* 2003, 13:2651-2657.

13. Jordan IK, Wolf YI, Koonin EV: **Correction: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly**. *BMC Evolutionary Biology* 2003, 3:5.
14. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al*.: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae***. *Nature* 2000, 403:623-627.
15. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: **Precision and functional specificity in mRNA decay.** *Proc Natl Acad Sci U.S.A.* 2002, 99:5860-5865.
16. 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.
17. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, *et al*.: **Widespread aneuploidy revealed by DNA microarray expression profiling.** *Nature Genetics* 2000, 25:333-337.

Table 1. Quantile analysis of mass spectrometry protein interaction data. Genes were separated into 1-5 bins based on their expression levels [15]. Each column is an analysis of the data set with a different number of bins. Each row is the rank of each bin's average expression level, where low rank indicates low expression. The upper number in each cell is the Spearman rank correlation coefficient between number of interactions and evolutionary rate (r_{EI}). The lower number in each cell is the partial correlation coefficient between number of interactions and evolutionary rate, controlling for expression level ($r_{\mathrm{EI.A}}$); the three cases in which this number has a greater absolute value than r_{EI} are underlined. *, p<0.05; **, p<0.005; ***, p<0.0005.

	1	2	3	4	5
1	–0.403*** –0.277***	–0.223*** –0.216***	–0.182* –0.182*	–0.179* –0.179*	–0.206* –0.205*
2		–0.341*** –0.285***	–0.323*** –0.328***	–0.245** –0.272**	–0.192* –0.197*
3			–0.263*** –0.258**	–0.366*** –0.358***	–0.352*** –0.336**
4				–0.225* –0.222*	–0.357*** –0.325**
5					–0.253* –0.207*