Heredity (2004) 93, 523–524

www.nature.com/hdy

NEWS AND COMMENTARY

Evolutionary proteomics

Computer dating for proteins

N Cobbe and MMS Heck

Heredity (2004) **93,** 523–524. doi:10.1038/sj.hdy.6800594
Published online 27 October 2004

To truly understand a protein's function in either a cellular or developmental context requires an appreciation of the other proteins with which it interacts. A broad range of experimental techniques have been developed to address this aspect of protein behaviour, ranging from yeast two hybrid assays and classical genetic screens with available mutants, to affinity chromatography, coimmunoprecipitation, protein crosslinking and phage display. However, no one technique is either infallibly accurate or sufficiently sensitive to detect all true physiological interactors. Consequently, most discerning biologists will not usually be convinced of the significance of a protein interaction either *in vitro* or *in vivo* unless strong evidence can be produced using two or more independent approaches. As most available screening methods are also relatively labour intensive, it is of considerable interest to see if the growing wealth of evolutionary data emerging from our present genomics era can be exploited in computational searches for interacting proteins. A number of methods have previously been employed to identify protein interactions based primarily on sequence data, including phylogenetic profiling, defined by the simultaneous presence or absence of certain proteins in different genomes (Pellegrini *et al*, 1999), and the identification of proteins in some organisms in which the domain structure consists of fusions between separate corresponding proteins found in other species (Enright *et al*, 1999). In prokaryotes, possible physical interactions between proteins may be inferred from conserved gene pairs in operons. However, each of these prediction methods is limited in its applicability, so more general approaches to identify candidate protein–protein interactions have been developed based on correlated evolutionary constraints, as discussed in a recent paper by Fraser *et al* (2004).

The notion that interacting proteins might evolve at similar rates can be understood in terms of complementary substitutions to maintain associations between them. This may involve either amino-acid residues that directly form intermolecular contacts or the indirect effects of neighbouring residues on domain conformation. Whereas mutations that perturb interactions would usually be removed by selection, this might be prevented by compensatory substitutions in interaction partners. Additionally, proteins that have many interactors generally evolve slowly as a greater proportion of their total length may be involved in functional interactions (Fraser *et al*, 2002). Although coevolution between interacting proteins may be inferred when proteins are simultaneously lost in the same species, specific interactions may also be detected based on statistically significant similarities between phylogenetic trees of paralogues (Pazos and Valencia, 2001). For example, the coevolution of a family of protein ligands and their receptors has been exploited to identify interacting pairs based on correlations between their phylogenetic distance matrices (Goh and Cohen, 2002). Correlated rates of substitution have also been demonstrated between regions of proteins that are either known or expected to interact with each other on the basis of biochemical or microscopy data (Cobbe and Heck, 2004), while empirical protein–protein interaction data have been used to deduce probable domain–domain interactions from which additional, previously unknown protein–protein interactions could then be inferred (Deng *et al*, 2002).

Previously, Fraser and Hirsh had shown that coevolution is probably the most likely explanation for observed similarities in evolutionary rates between interacting proteins in yeast, rather than alternative explanations such as similar fitness effects or overall numbers of interactors (Fraser *et al*, 2002). In their recent paper (Fraser *et al*, 2004), these authors now demonstrate how independent evidence for protein coevolution may be provided by not just examining similarities in evolutionary rates but also evaluating the extent to which genes may resemble each other in their presumed average expression levels. Firstly, evolutionary distance was estimated by calculating the average ratio of amino-acid replacement or nonsynonymous nucleotide substitutions to silent or synonymous substitutions at each codon position of orthologues from four closely related yeast species of the genus *Saccharomyces*. As well as indicating the divergence between proteins, this ratio also provides a measure of the intensity of purifying or negative selection exerted on genes, which results in decreased rates of amino-acid replacement and thus substitution ratios less than one. Consequently, any observed correlations should also reflect sequence coevolution based on adaptive changes, rather than just similarities in divergence patterns. Using normalised values for orthologues displaying greater variance in their rates, it was found that more significant correlations were observed between the evolutionary rates of protein pairs known to interact than with random pairs, in agreement with previous studies.

The extent of coevolution in gene expression was then examined, as previous studies have shown that functionally related or interacting proteins have a significantly higher correlation between their gene expression profiles than random protein pairs. Analysis of mRNA coexpression has also been used to detect functional relationships between proteins in combination with other evidence such as phylogenetic profiles or domain fusions (Marcotte *et al*, 1999). However, rather than using expression profiles derived from DNA microarrays, Fraser *et al* employed patterns of codon usage to estimate the average expression level of genes in the same four yeast species. In such species, a strong association between codon bias and expression levels is thought to reflect selection for increased translational efficiency of highly expressed genes (corresponding to the most abundant tRNA in a given species) in addition to requirements for increased translational accuracy of essential proteins. Moreover, codon bias relationships should provide a more robust indication of evolutionarily conserved trends in gene expression, as codon usage is heritably selected but expression profiles may be influenced by experimental conditions. In the current paper, the extent to which codon usage patterns resemble those of highly expressed genes was measured by calculating the codon adaptation index for each gene, a statistic indicating the extent to which the pattern of codon usage resembles that of genes known to

be highly expressed. In this way, it was found that the majority of interacting proteins displayed dramatically higher correlation coefficients in their levels of expression than previously observed with evolutionary distances. Importantly, the authors also ruled out the alternative possibility that observed correlations between codon usage patterns might be due to coregulation by the same transcription factors, since significantly lower correlations were observed between most coregulated genes. It therefore appears that coevolution of expression may be an even more powerful predictor of physical interaction than sequence coevolution. However, the correlation between both of these measures of coevolution was found to be extremely weak, indicating their apparent independence. More accurate predictions of physical interactions could therefore be obtained by combining these two measures, simultaneously evaluating coevolution of amino-acid sequence and gene expression. Strikingly, a 27-fold enrichment for interacting proteins (compared to randomly selected yeast proteins) was obtained by combining both measures of coevolution, with different arbitrary cutoffs for correlation between estimated expression levels or evolutionary distance.

In the future, it will be interesting to determine if the correlations between evolutionary rates in whole proteins may be further refined in subsequent attempts to identify probable interacting domains (Deng *et al*, 2002; Cobbe and Heck, 2004). It is also worth considering whether generalised measures of codon usage bias (such as the effective number of codons) or correlated distance measures reflecting differences in all codon frequencies might also facilitate the identification of functionally related proteins in organisms for which reliable optimal codon data are not yet available. Although the present study has focused on budding yeast genomes, it is expected that the same approach will be applicable to other genera, such as *Drosophila* and *Caenorhabditis*, in which biased codon usage is clearly correlated with gene expression levels (Duret and Mouchiroud, 1999) and genome sequences are available for more than one species. On the other hand, the accumulating data generated by microarray-based analyses of gene expression should also facilitate the identification of potential interactors under specific developmental conditions in a host of additional species. These data may be complemented by analyses of coevolution in substitution rates and gene expression that reflect coordinated changes in the use of distinct developmental, metabolic or behavioural pathways. Consequently, the same combined methodology may be useful both in verifying significant protein interactions from high-throughput data sets and identifying novel interactors in other organisms based on both similar evolutionary rates and similar expression levels.

N Cobbe and MMS Heck are at the The Wellcome Trust Centre for Cell Biology, Institute of Cell Biology, University of Edinburgh, Scotland EH9 3JR, UK.

e-mail: Neville.Cobbe@ed.ac.uk

Cobbe N, Heck MMS (2004). *Mol Biol Evol* **21**: 332–347.
Deng M, Mehta S, Sun F, Chen T (2002). *Genome Res* **12**: 1540–1548.
Duret L, Mouchiroud D (1999). *Proc Natl Acad Sci USA* **96**: 4482–4487.
Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999). *Nature* **402**: 86–90.
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002). *Science* **296**: 750–752.
Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004). *Proc Natl Acad Sci USA* **101**: 9033–9038.
Goh CS, Cohen FE (2002). *J Mol Biol* **324**: 177–192.
Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999). *Nature* **402**: 83–86.
Pazos F, Valencia A (2001). *Protein Eng* **14**: 609–614.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999). *Proc Natl Acad Sci USA* **96**: 4285–4288.