The human protein coevolution network

Elisabeth R.M. Tillier[1] and Robert L. Charlebois

Department of Medical Biophysics, University of Toronto, and Ontario Cancer Institute, University Health Network, Toronto, Ontario M5G 1L7, Canada

Coevolution maintains interactions between phenotypic traits through the process of reciprocal natural selection. Detecting molecular coevolution can expose functional interactions between molecules in the cell, generating insights into biological processes, pathways, and the networks of interactions important for cellular function. Prediction of interaction partners from different protein families exploits the property that interacting proteins can follow similar patterns and relative rates of evolution. Current methods for detecting coevolution based on the similarity of phylogenetic trees or evolutionary distance matrices have, however, been limited by requiring coevolution over the entire evolutionary history considered and are inaccurate in the presence of paralogous copies. We present a novel method for determining coevolving protein partners by finding the largest common submatrix in a given pair of distance matrices, with the size of the largest common submatrix measuring the strength of coevolution. This approach permits us to consider matrices of different size and scale, to find lineage-specific coevolution, and to predict multiple interaction partners. We used MatrixMatchMaker to predict protein–protein interactions in the human genome. We show that proteins that are known to interact physically are more strongly coevolving than proteins that simply belong to the same biochemical pathway. The human coevolution network is highly connected, suggesting many more protein–protein interactions than are currently known from high-throughput and other experimental evidence. These most strongly coevolving proteins suggest interactions that have been maintained over long periods of evolutionary time, and that are thus likely to be of fundamental importance to cellular function.

[Supplemental material is available online at http://www.genome.org. MatrixMatchMaker is freely available at http://www.uhnres.utoronto.ca/labs/tillier/MMMWEB/MMMWEB.php.]

Cellular proteins rarely work in isolation, but are more usually involved in pathways and interaction networks. Dysfunctional networks are commonly implicated in disease, so an elucidation of protein–protein interactions can greatly contribute to our understanding of pathological states, and much more broadly, of molecular biology. Systems biology relies on accurate representations of interaction networks, but these are often hard to describe. Interactions can be conditional or contextual, and may not always be captured in a given study, regardless of its attention to quality. Complementary approaches based on experimental data as well as on sequence and evolutionary analyses are required in order to describe a system with a sufficient degree of detail so that it may effectively be understood.

Large protein–protein interaction networks are commonly obtained from high-throughput experiments. Such large-scale efforts have described the interaction maps for organisms such as *Helicobacter pylori* (Rain et al. 2001), *Escherichia coli* (Butland et al. 2005), *Saccharomyces cerevisiae* (Ito et al. 2000; Schwikowski et al. 2000; Uetz et al. 2000; Gavin et al. 2002; Ho et al. 2002), *Caenorhabditis elegans* (Walhout et al. 2000; Simonis et al. 2009), *Drosophila melanogaster* (Giot et al. 2003), and *Homo sapiens* (Rual et al. 2005; Stelzl et al. 2005). These studies have been extremely valuable toward our understanding of protein interaction networks, but they suffer from inherent experimental biases in terms of the types of interaction that can be detected. The first *S. cerevisiae* network, for example, was estimated to have a false-negative rate of 90% and a false-positive rate of 50% (von Mering et al. 2002; Sprinzak et al. 2003). Nevertheless, accuracy can be increased by combining data sets (Bader and Hogue 2002; von Mering et al. 2002; Yu et al. 2008), by repeated screening (Venkatesan et al. 2009) and confidence evaluation (Yu et al. 2008; Braun et al. 2009), and by comparing data sets from different species (Kalaev et al. 2008). Several underlying principles of protein interaction networks have thus been discovered. Their scale-free nature (Barabási and Oltvai 2004) suggests that there are a small number of very highly connected proteins (hubs), which form highly connected modules of proteins sharing well-defined functions (Snel et al. 2002; Rives and Galitski 2003; Spirin and Mirny 2003; Wuchty and Almaas 2005). These hub proteins tend to be found in more species, to have higher sequence conservation and thus to be slow to evolve, and are essential for survival (Fraser et al. 2002, 2003; Jordan et al. 2003; Wuchty et al. 2006). Although many protein interactions appear to be lineage-specific (Mika and Rost 2006; Cusick et al. 2009), evolutionarily conserved proteins maintain their interaction partners (Wuchty et al. 2006; Yellaboina et al. 2008).

Many examples of interacting protein families have been shown to coevolve, with interacting members of the different families displaying similar phylogenetic trees (Moyle et al. 1994; van Kesteren et al. 1996; Goh et al. 2000). This observed correlation between the pattern and rate of evolution of members of functionally related protein families is thought to be the result of the functional complementarity of gene duplication and divergence events (Fryxell 1996). Kim et al. (2004) analyzed protein family pairs and found significant correlation in the evolutionary patterns for 78% of the 454 that were reliably known to interact. Hakes et al. (2007) showed that the correlation of evolutionary rates between interacting proteins in yeast cannot simply be explained by the covariation between functional residues in isolation. However, others have been able to predict interacting domains from coevolving residues between domains or proteins (Jothi et al. 2006; Kann et al. 2007, 2009; Yeang and Haussler 2007).

[1]Corresponding author.
E-mail e.tillier@utoronto.ca; fax (416) 581-7581.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.092452.109.

Great efforts have been made in using sequence information alone to predict interacting proteins and domains (for review, see Shoemaker and Panchenko 2007; Jothi and Przytycka 2008), all based to some degree on coevolutionary arguments. One method has been to consider the conservation of gene order, whereby an interaction influences the relative placement of these genes in the genome, in some cases fusing the genes altogether. Another has been to consider the phylogenetic profile of genes, whereby the presence or absence of a gene influences the presence or absence of other genes.

Coevolution, as more directly measured by the similarity of evolutionary trees, is often used to predict interacting proteins and domains (Ramani and Marcotte 2003; Valencia and Pazos 2003; Pazos et al. 2008). The technique does not normally compare trees directly, but rather through their implied distance matrices (although see Jothi et al. 2005 for a tree-based approach). Correlated values between two matrices reflect similar evolutionary trees, which can indicate coevolution, although as with amino acid covariation, the difficulty arises in separating the background signal of functional coevolution from the species phylogeny (Pazos et al. 2005; Sato et al. 2005). Recently, Juan et al. (2008) extended the mirror tree approach for pairwise observations toward a complete cellular "coevolutionary context." This was achieved by considering not only the individual correlations between pairs of proteins in the *E. coli* proteome, but also the secondary correlations with the entire set of proteins (a technique previously described by Sato et al. 2006). By using this coevolutionary context information, they predicted the interactome of *E. coli* with a degree of accuracy and coverage comparable with that of the high-throughput experimental techniques.

The mirror tree approach for detecting coevolution between proteins identifies protein partners by maximizing the correlation between the distance matrices describing a pair of phylogenies. Our own earlier program Codep (Tillier et al. 2006) finds the mutual information between all of the columns from two alignments to determine pairs of proteins that maximize this proxy for the coevolutionary signal. Tree-based and coevolutionary methods can detect interactions between genes and have worked particularly well for proteins from the smaller, compact bacterial genomes, for which many orthologous sequences are available for comparison, and where unique copies of protein-coding genes in genomes can be assessed for coevolution with other such unique genes. In situations in which there has been gene duplication, binding specificities of paralogous proteins may diverge after duplication, and there may be concerted duplications of the interaction partners. Particular members of a gene family within a genome may then have a particular affinity to select members of another protein family. However, the combinatorial problem raised by the presence of paralogs and the potential for multiple interacting partners make predicting protein interactions using coevolution in large genomes with many paralogs difficult. Another complication arises owing to gene loss, either from deletion or through high sequence divergence, which can make identifying orthologous copies and accurately measuring divergence also difficult. Both gene duplication and gene loss will give rise to protein families of different sizes, which remains outside the scope of these bioinformatics methods.

Here we describe a novel approach for measuring the coevolution of protein families that address many of the drawbacks of previous methods. Our approach, that we call MatrixMatchMaker (MMM), finds the largest common submatrix that is compatible between the evolutionary distance matrices of two protein families. It can measure the coevolution between families of different size, including both paralogs and orthologs, and predict multiple coevolving partners. The method allows us to detect coevolution present in only some lineages, thus suggesting lineage-specific interactions. Here, we use MMM to analyze the human protein interaction network, to show that a large fraction of known protein interactions, as well as proteins belonging to the same biochemical pathway, do indeed coevolve.

MMM, applied to the analysis of protein families, proposes several novel interactions. Although only a subset of coevolving proteins would be expected to interact, and vice versa, the predictions afforded by the use of the present analysis provide a highly enriched set of hypotheses that can more efficiently direct laboratory validation. The human protein coevolution network is large, but can be focused to cover a more restrictive scope. Where proteins are involved in a tight coevolutionary relationship implying a strict interdependency, we may find some of the more fundamental interactions, core to the system in which these proteins are engaged.

Results

A new approach to measuring coevolution: MMM

MatrixMatchMaker (MMM) provides a novel strategy for studying the coevolution of proteins. It uses a bottom-up strategy that seeks only those sequences most strongly implicated in a coevolutionary relationship. It therefore promises to be more flexible in its assumptions about coevolution and much more convenient in its accommodation of the data at hand. Matrices can contain various numbers of sequences evolving along independent trajectories and can even include irrelevant sequences perhaps erroneously included in a multiple sequence alignment.

MMM implements a strategy to detect sequence coevolution from a pair of distance matrices (typically of different size) by finding their largest common submatrix. We avoided comparing phylogenetic trees directly because of their inherent compromises and topological inconsistencies (Waddell et al. 2007), but conceptually, the submatrix returned by the program represents the largest common subtree. A submatrix match is built up by sampling each distance matrix systematically, in order to find entries that can be coordinately added to the growing match with a desired consistency of fit (see Methods for details). Since the two matrices may be scaled differently, we use relative distances in judging matches, and a match is deemed satisfactory when it does not exceed a user-supplied tolerance threshold, MMM's main parameter. The more stringent this parameter, the stronger the similarity must be in comparing relative distances, and therefore the stronger the coevolutionary signal between the proteins. Since several different largest common submatrices might be found, MMM returns all solutions (of minimum size three) ranked by score in a file that lists the correspondence between sequence labels.

For this study, we used an option offered by MMM that allows sequences to be annotated with an additional taxon label and requires these taxon labels to match (e.g., matching human proteins only to human proteins and yeast to yeast). We furthermore only considered submatrices that included at least one human protein, requiring coevolution to be observed in the human lineage. Using these restrictions, appropriate to our purpose, no computational impediments were observed in analyzing the 6 million pairs of available matrices.

Using known interactions to validate and calibrate MMM

The HomoloGene database (Wheeler et al. 2008) was used as our source of sequence information as it includes protein sequences from 20 complete eukaryotic genomes that are clustered by homology (including paralogs and orthologs). To validate our approach, estimate its accuracy, and determine optimal parameters for the MMM algorithm, we considered the set of human "known" protein interactions from protein interaction databases (NCBI Gene [Wheeler et al. 2008], BioGRID [Breitkreutz et al. 2008], and HPRD [Peri et al. 2003; Mishra et al. 2006]) and used HomoloGene's clusters to extrapolate interaction data known from other species to human (interologs). Mapping physical interactions using interologs is controversial, as it has been shown that physical protein interactions are more conserved within species than they are across species (Mika and Rost 2006). However, we considered that the coevolution of proteins would supply additional evidence for interaction above simple conservation. We also considered the KEGG database (Kanehisa et al. 2004) to obtain pathway information and thus to relate HomoloGene clusters in terms of shared pathways. The networks were then annotated with additional information from the Gene Ontology (GO) database (Ashburner et al. 2000). HPRD can be considered a source of high-quality interactions, but since these curated data are few, we optimized our analysis to maximize the accuracy of MMM relative to all of the "known" interactions, assuming that these interactions are real despite a high degree of inaccuracy and lack of overlap known among the methods (both high and low throughput) that populated the databases.

Coevolution attributed to common phylogenetic diversity

The MMM score shows remarkable separation of known interacting proteins from the rest (Fig. 1A). Most of the signal can be explained, however, by the fact that interacting proteins are more prevalent among the species included and have followed a common species phylogeny. Proteins that follow the main phylogenetic signal and are conserved enough to be found in many species will give larger distance matrices that are more compatible with other matrices. To demonstrate this, for each pair of matrices we calculated the match potential between them, which measures the compatibility of two matrices based on the number of sequences in taxa in common. Interacting proteins have higher match potential (Fig. 1B). These results confirm a previous study of the yeast interactome, which also showed a propensity of interacting proteins to be more conserved and prevalent among taxa (Wuchty et al. 2006). In a way, this is a form of phylogenetic profiling, analogous to that used for bacterial genomes (Pellegrini et al. 1999), but here applied to eukaryotes (also see Cusick et al. 2009).

Coevolution attributed to correlated rates of evolution

Although much of the general coevolutionary signal might be attributable to prevalence and phylogeny, we would expect coevolution among interacting proteins to be further constrained by the nature of interaction itself, and thus for coevolution of interacting proteins to be strongest. To determine if interacting proteins are more tightly correlated in their rates of evolution, we ran MMM with increasing stringency. This progressively removes pairs of proteins with the more marginal coevolutionary signals. We found that making the tolerance more stringent increased the relative frequency of interacting protein pairs found by MMM, indicating that these have more correlated rates of evolution over longer evolutionary time (Fig. 2). This clearly demonstrates that coevolution, acting at the level of correlated evolutionary rates, provides predictive power for MMM to discover interacting proteins.

Quality of distance matrices for coevolutionary analysis

We expected the accuracy of the method to be highly dependent on the quality of the distance matrices. We performed the MMM analysis with distance matrices calculated using ClustalW (Thompson et al. 1994), but also with matrices calculated using a corrected distance measure using Protdist from PHYLIP (Felsenstein 1989) with a PMB substitution model (which most resembles the BLOSUM model used by ClustalW; Veerassamy et al. 2003). We found that corrected distances from Protdist retrieved smaller compatible matrices than did ClustalW (Fig. 1). This makes sense, as corrected distances increase exponentially, and thus differences tend to be magnified as distances grow large and it becomes more difficult to satisfy MMM's match criteria. ClustalW-type distances are thus better suited for coevolutionary analysis as had previously been observed (Tillier et al. 2006; Izarzugaza et al. 2008). However, ClustalW distances can present a very high probability of false-positive identification when sequences are saturated with substitutions and the alignment becomes inaccurate. Highly diverged sequences converge to the maximum distance of 1, such that in contrast to corrected distances, they become more similar.

To circumvent these properties of distance matrices, we used ClustalW distances, but also implemented a stepwise linear function for the tolerance parameter in MMM (see Methods for details) that allowed us to increase the match stringency for very high (converging) or very low (potentially noisy) distances. Using this function also allows us to ignore sequences with very poor alignments to the rest of the sequences

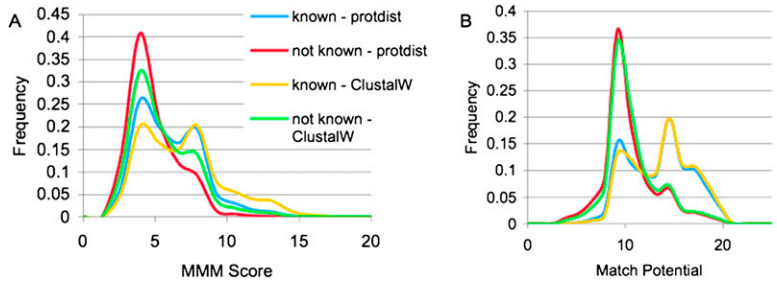

Figure 1. Known physical interactions have larger MMM scores. (*A*) The frequency distribution of MMM scores is shifted to higher values for known physically interacting proteins (blue and yellow) when compared to all matrix pairs for which no physical interaction is reported in the databases (red and green). This indicates stronger coevolution for known interacting proteins, and the difference is more pronounced when using matrices calculated with ClustalW (yellow vs. green) than with Protdist (blue vs. red). (*B*) Frequency distributions of the match potential, which is simply the largest possible match size theoretically obtainable in a comparison of the two matrices. The higher values seen here for known interactions indicate that these proteins are more prevalent and conserved. These plots were obtained with the MMM tolerance parameter set to 0.2 (allowance for 20% relative length mismatch) and without the use of a stepwise linear function adjusting this tolerance for extreme values of distance.

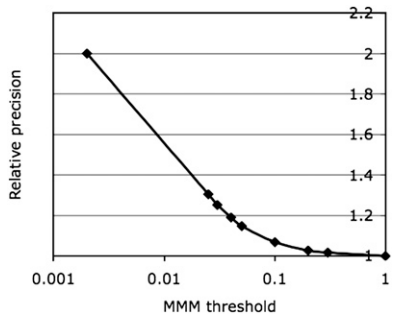

Figure 2. The accuracy of calling true positive (known physical) interactions increases with a lowering of the MMM tolerance parameter threshold. To generate this figure, MMM was run with decreasing tolerance from 1.0 to 0.0025. The accuracy is normalized to that obtained when using the most generous tolerance of 1.0.

in a HomoloGene cluster, sequences that might not actually belong in those clusters.

Comparison with the matrix correlation method

Classic matrix comparison methods correlate entire distance matrices, necessarily of equal size, in order to evaluate whether their component sequences are coevolving (Ramani and Marcotte 2003; Valencia and Pazos 2003; Sato et al. 2005; Pazos et al. 2008). Preprocessing of matrices is often required to ensure mutual correspondence of the same number of homologs. Because of this constraint, such methods cannot consider multiple unequal numbers of paralogs and can neither consider coevolution in a subgroup of species where one of the proteins is absent. If only a subset of sequences from each matrix is coevolving, as would happen when interacting gene families diversify and specialize their partnerships independently, then a whole-matrix approach could lack the necessary sensitivity.

We compared MMM to a basic correlation approach. First, we identified the 106,179 pairs of matrices from the more than 6 million directly suitable for MMM analysis that had identical size and taxon distribution in their HomoloGene clusters. This gave us a large enough sample of matrix pairs on which we could directly compare the two methods. The Pearson correlation coefficient was calculated between each of these matrix pairs. Because of the strong phylogenetic component to the correlation, and more trivially because of the simultaneous presence of both small and large distances in many matrices, these values tended to be quite high. In Figure 3A, we show the distribution of correlation measures for known interactions (from HPRD, BioGRID, NCBI Gene, and KEGG: 895 interactions in total), the best quality interactions (85 only, from HPRD), and pairs for which no interaction evidence was found among these databases. Strong correlation does not appear to be predictive of known interactions in this data set, as there is little separation between the distributions for known interacting proteins from the rest, even when considering the highest-quality interactions from HPRD. In Figure 3B, however, we see a clear overrepresentation of known interactions for matrices with high MMM scores. When we correct the Pearson correlation for phylogenetic signal (see Methods), the performance improves, as is shown using a ROC curve analysis (Fig. 3C). However, we find that the MMM score is still a better predictor of known interactions in this data set.

This result demonstrates that it is important to consider tree size and covariation over different parts of the tree as is done by MMM, rather than just an overall correlation as is done with the matrix correlation approach. In Supplemental Figure 1, we show the relationship between the Pearson correlations and MMM scores for these matrices. Although high MMM scores require a strong correlation between the matrices, the reverse does not hold, as strongly correlated matrices can have low MMM scores. With its constraints on a priori sequence correspondence and with its sensitivity to correlation artifacts, the matrix correlation method may not be as generally suitable as MMM, at least for the type of problem here at hand.

Analysis of physical interactions and KEGG pathways with MMM

As described above, many interacting proteins coevolve, such that we can use MMM to predict interacting proteins in this eukaryotic data set. To obtain a coevolution network, several parameters must be considered. These are the threshold for the tolerance of differences in the relative distances and the values describing the stepwise linear function correcting for limitations in the range of distances. We can also consider another threshold, requiring a minimum size for the compatible submatrices. To predict

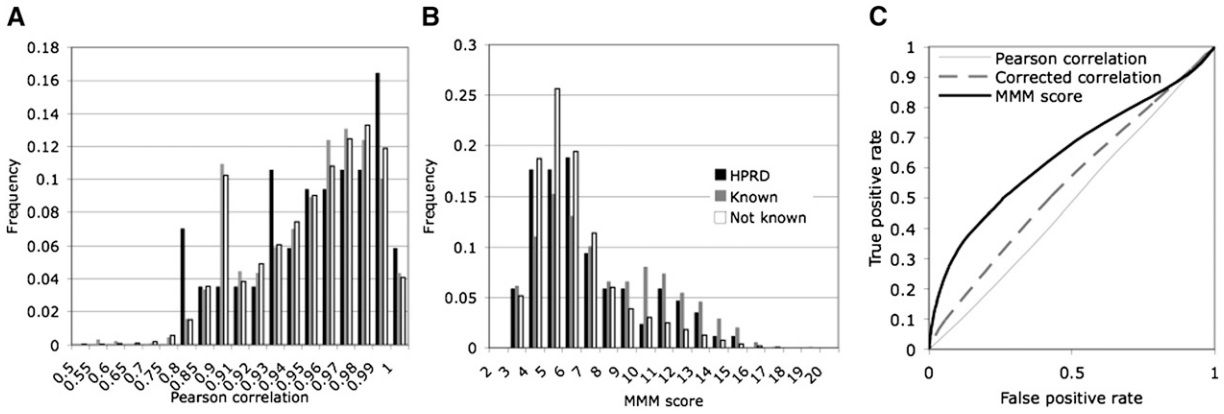

Figure 3. Comparison of matrix correlation scores and MMM scores. (*A*) The distribution of the frequency of known interactions for different degrees of Pearson correlation coefficient is plotted for the 106,179 matrix pairs with equal size and taxonomic distribution. The Pearson correlation scores are very high and do not segregate known interactions, nor high-quality interactions (HPRD). Key shown in *B* is the same as for *A*. (*B*) We do, however, see a clear excess of known and HPRD interactions with high MMM scores. A ROC curve shows that the accuracy of the Pearson correlation is improved when corrected for phylogenetic signal, and that the MMM score is still a better predictor. MMM was run with a tolerance of 0.2 and no step function.

interacting proteins, we sought to find protein pairs that have large compatible distance matrices, but that are also not so conserved as to have insufficient evolutionary signal and that are not so diverged that the difference in their sequences is saturated. We constructed several interaction networks obtained using different combinations of MMM thresholds, a sampling of which is shown in Supplemental Table 1.

The maximum number of known physical interactions for 3471 gene families in the analysis was 19,854, or 43,493 if KEGG pathways were also included. The total pairwise number of combinations of matrices was more than 6 million, such that known interactions consist of only 0.33% (0.72% with KEGG) of potential interactions. We saw that when using MMM at stricter thresholds, there is a clear enrichment for known interactions (Fig. 2), again illustrating that known protein interactions are among the most highly coevolving proteins. We also see that coevolution more strongly predicts physical interaction than functional interaction.

Since we are considering potentially very noisy interaction data in our set of previously known interactions, it was important to determine the significance of the number of these interactions in the coevolution network. We determined the statistical significance by randomizing the gene names included in the network, but otherwise keeping the network structures the same. We found all networks obtained with the given set of clusters to be highly significant (by Z-score, with P-values $< 10^{-14}$).

Besides the MMM score, additional considerations can be applied to try to further increase the accuracy of this coevolutionary analysis, so we explored the effect of removing the overall phylogenetic signal and of strengthening network ties by correlating scores (see Methods). A plot showing the accuracy (true positives over all positives, also called "precision") over the full range of thresholds of different MMM-generated scores is shown in Supplemental Figure 2.

Accuracy as measured by precision does not consider the trade-off that, as we increase the frequency of true positives, we are also reducing the number of total positives. To further analyze our results, we considered the F_β-score, which is a mean of precision and recall (van Rijsbergen 1979) given by

$$F_\beta = \frac{(1+\beta^2)\{precision\}.\{recall\}}{\beta^2\{precision\}+\{recall\}}.$$

Precision is the number of true positives divided by the total number of positives. Recall is defined as the number of true positives divided by the sum of true positives and false negatives. Since we are still much more interested in the accuracy of MMM in terms of correctly identifying known interactions (precision), rather than identifying the most known interactions (recall), we used a low β of 0.1 such that precision is weighted 10 times more than recall. Supplemental Figure 3 shows the maximum $F_{0.1}$ score obtained when the full range of possible thresholds was applied to several different scores, for several different runs of MMM with different parameters.

The accuracy when considering KEGG pathways is approximately the same as when considering physical interactions, but these are in most cases complementary as overall accuracy increased substantially when considering both. Physical interactions yielded larger coevolving trees (Fig. 4B), indicating a more consistent degree of coevolution than from mere pathway membership. Because a large proportion of known physical interactions come from yeast, we considered that this may have caused a bias in favor of large matrices, since matrices containing conserved proteins from yeast to human should be overrepresented. KEGG pathways represent more general biochemical knowledge and should thus be expected to be less prone to such bias. However, there is no evidence for such a bias, as we could not find a difference between the match potential of matrices with known physical interactions and those in the same KEGG pathways (Fig. 4A). Belonging to a common pathway does promote coevolution, but not as strongly as does physical interaction. Subtracting the average matrix—the phylogenetic signal—decreases the accuracy for KEGG pathways, suggesting that their coevolution is largely directed by the overall phylogeny (Fig. 4C). In contrast, this correction ameliorates the accuracy for proteins involved in physical interactions, implying that their coevolution is in part contrary to phylogenetic signal, and that one protein's evolution drives another's. Still, many physically interacting proteins do coevolve along the tree.

KEGG interactions have more strongly correlated MMM scores (Fig. 4D), indicating larger networks for these proteins. Considering the MMM score reduced by the maximum best score of either matrix in the pair (score E in Methods) can give an indication of the specificity of the predicted interaction, since a low score would indicate that there is at least one other pairing for the protein with an equivalent MMM score. Physical interactions are

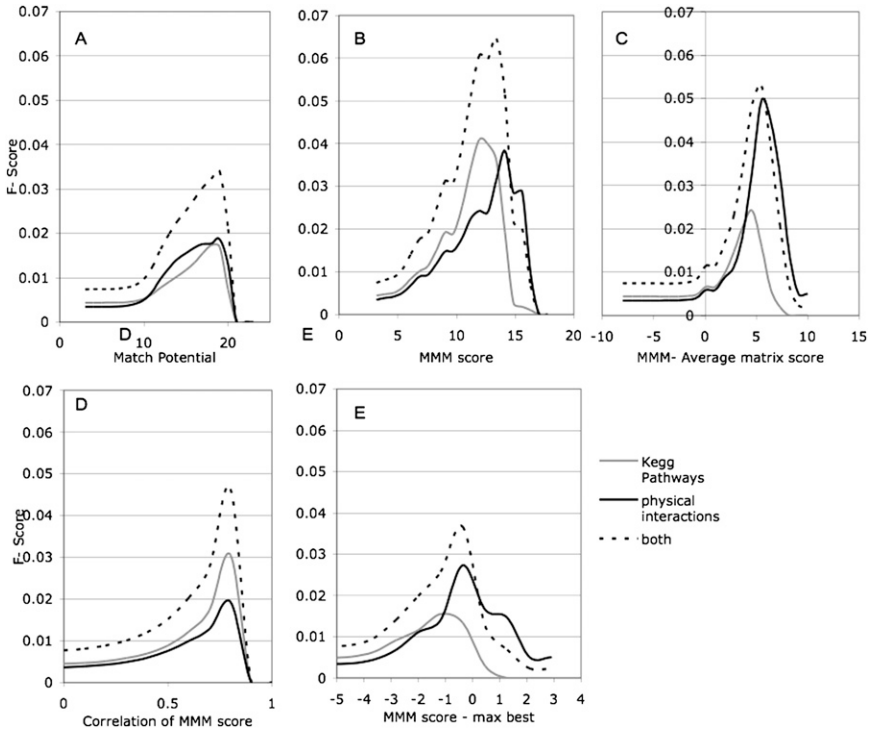

Figure 4. Comparison of $F_{0.1}$ scores for KEGG and for physical interactions. MMM was run with parameters -a 0.2 -s1 0.02 -s2 0.5 -s3 0.9 (see Methods for a description of the various scores).

much more likely to be specific to the pair, compared to KEGG interactions (Fig. 4E). One would expect physical interactions to be more constrained, where cascades of codependent sequence changes should occur, and for this coevolution to be fairly specific to a pair of proteins.

Prediction of the human protein coevolution network

We first considered the small network of interactions for the most strongly coevolving genes in our data set (i.e., with our most stringent MMM parameters) (Fig. 5; Supplemental Cytoscape file, Shannon et al. 2003). A total of 184 genes are included in this network, with 156 edges, 11 of which are known physical interactions (red edges) with an additional 12 being annotated to the same KEGG pathway (blue edges).

We see several protein complexes in this coevolution network involved in fundamental cellular processes of protein synthesis and cell division. For example, we find many aminoacyl tRNA synthetases (AARSs) that are known to have additional noncanonical activities beyond their role of linking specific amino acids to tRNAs (Park et al. 2005). Some AARSs have been shown to form high-molecular-weight complexes in nuclei, suggesting they could play a role in DNA replication as well (Nathanson and Deutscher 2000). We find coevolution with members of the minichromosome maintenance complex (MCMC), proteins involved in replication as a part of the pre-replication complex. MCMC proteins have also been implicated in diverse chromosome transactions including genome stability, transcription, and chromatin modification (Tabancay and Forsburg 2006). We find some of these MCMC proteins to be coevolving with TCP1 and other members of the CCT complex. The CCT complex is a protein chaperone thought to fold actin, tubulin, and other newly synthesized polypeptides (Thulasiraman et al. 1999). Although usually cytoplasmic, CCT has been shown to enter the nucleus and to be associated with constitutive heterochromatin and to be involved in compacting chromatin, raising the possibility that it may be implicated in maintenance and remodeling of heterochromatin in mammalian spermatogenesis (Souès et al. 2003). A recent study, using proteomic and genomic approaches, considered the interactions of the CCT complex in the yeast genome (Dekker et al. 2008). Although interactions with MCMC proteins were not found, they did find interactions with many proteins involved in chromatin remodeling. The links around *TCP1* in our network are strong even when corrected for the overall phylogenetic signal (that score determines the width of the edges in Fig. 5).

Intriguingly, we also find strong links for the PKM2 pyruvate kinase with AARSs and the CCT complex. In mammals, pyruvate kinase (PK) exists in the form of four isozymes, M1 and M2 transcribed from one gene and L and R from another. These are differentially expressed in different cell types (Tanaka et al. 1967; Jurica et al. 1998). The M1-type PK is well known for its role in glycolysis but has also been reported to influence microtubule stability (Vértessy et al. 1999). Confusingly, the sequence of the *PKM2* gene in the HomoloGene cluster is that of the M1 type. (The M2 isoform, a result of differential splicing, contains only 21 amino acid differences with M1. M2's sequence is not included in the HomoloGene cluster, and whether PK M2 reveals the same coevolutionary partners as PK M1 is left for another study.) The M2 isoform, normally thought to be embryonically restricted, is expressed in cancerous cells (Altenberg and Greulich 2004). PK M2 was recently shown to be a phosphotyrosine-binding protein, switching cellular metabolism to aerobic glycolysis (Christofk et al. 2008). It was also shown that PK M2 is translocated to the nucleus in response to apoptotic agents and that it induces programmed cell death (Steták et al. 2007).

These proteins in the strongest coevolving network are thus involved in the most fundamental cellular processes of DNA replication, protein synthesis, and energy metabolism. By reducing the stringency of MMM parameters by increasing the tolerance thresholds, we could obtain larger networks and thereby examine more general properties of coevolution networks. For example, we considered the larger MMM #5 network (given in the Supplemental material) and studied some of its network characteristics using the Network Workbench tool (http://nwb.slis.indiana.edu). We also used BiNGO (Maere et al. 2005) to find genes in the coevolution network whose GO-Full terms were significantly overrepresented (with respect to all of GO) and analyzed these in terms of the number of connections in the network (Supplemental Table 2).

Overall, this large coevolutionary network is overrepresented with respect

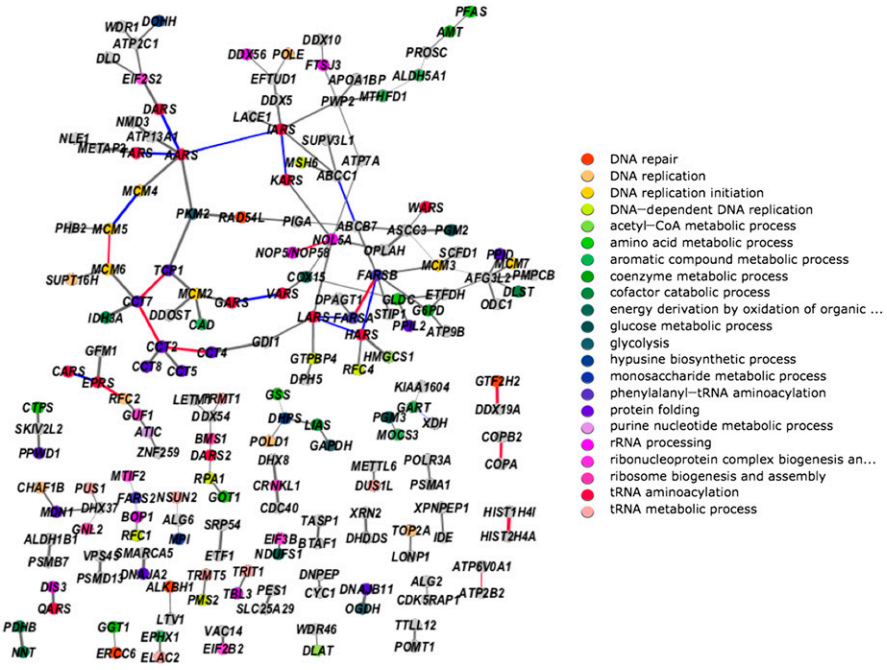

Figure 5. The coevolution network (#1 from Supplemental Table 1) of the most strongly coevolving proteins is shown with known physical protein interactions as red lines, and with KEGG pathway sharing as blue lines. The width of the lines is proportional to the MMM score minus the score to the average matrix. The proteins (nodes in the graph) are colored according to their GO annotation with the highest connectivity in the network. Given the detail in this figure, we also make it available as a supplemental Cytoscape file for more convenient viewing and manipulation.

to macromolecular complexes, nuclear and mitochondrial genes, and rRNA processing. We also find that this network is enriched for internal-membrane proteins and proteins involved in lipid metabolism, allowing us to make predictions for non-soluble proteins, which have been much less well characterized to date. The network is, however, underrepresented in cell communication, receptors, signal transduction, transcriptional regulation, cell mobility, and adhesion.

Discussion

Experimental studies cannot be counted on for identifying all of the interesting and relevant protein–protein interactions acting within a given system. Any high-throughput approach will produce its share of false-positive and false-negative identifications, whereas a more rigorously focused study must necessarily ignore the constellation of other proteins playing some role. In combination, many such studies together will contribute to a broader and deeper understanding of the intricate pathways and interaction networks that drive living systems. Bioinformatics approaches, such as MMM and others reviewed herein, have an important role to play as well. They can direct experiments and be informed by them, in an iterative cycle aiming toward a more complete and more robust understanding of the biological networks important in so many processes. Coevolutionary analysis provides insights into those components with mutual interdependencies that change in lockstep.

A prerequisite for predicting functional protein interaction using coevolutionary analysis is the availability of a sufficient amount of signal in the form of sequence divergence. From the breadth of eukaryotic genomes that we have considered here, coevolution appears to be readily measured in many cases, and such cases are enriched for known physical protein interactions as well as for proteins involved in the same biochemical pathway. Since these two sources of evidence for putative interaction improve detection accuracy cumulatively, additional examples of physical or biochemical interaction should further corroborate many of our predictions. We propose that coevolution can serve as a complementary approach to other methods to predict functional interaction networks. Although the majority of interacting proteins do not appear to coevolve, coevolving proteins remain an interesting subset.

The approach taken by MMM is more versatile than traditional matrix methods as it readily accommodates matrices of unequal size. This allows us to consider a more general model of coevolution with unequal duplication patterns and more recently derived protein interactions. Additionally, we found it to be more predictive for this eukaryotic data set. The combination of long and short branches between the species considered here automatically leads to highly correlated matrices, which would lead to false-positive identification of interactions. MMM uses a local search strategy to detect coevolution in that it requires submatrices to be correlated. It requires coevolution to occur throughout all branches, whereas traditional approaches only require an overall correlation and are more easily misled by the speciation signal.

Despite protein interactions often being mediated through interaction domains in eukaryotes, we still detected coevolution using full-length sequences. This could be a consequence of a strong phylogenetic signal tracking the pattern of speciation, and thus exaggerating false-positive identification. We found that a much higher proportion of known interacting proteins were uncovered if we did not attempt to remove the overall phylogenetic signal from the coevolutionary signal, suggesting that directed coevolution is at play. Uncorrected MMM scores gave the best accuracy, however, only when we sought the most highly coevolving proteins. By relaxing the match threshold, we could obtain very large networks, but at the expense of more false positives. Accuracy in these more expansive networks benefited from a correction for phylogenetic signal as well as from the use of correlated scores.

Correlating scores did not, however, produce as much of an increase in the accuracy of predicting protein interactions as had been found previously for the *E. coli* network (Juan et al. 2008). We used the correlations of scores from the MMM analysis on eukaryotic data, so it is difficult to compare our results directly; however, we think that the difference may be due to the nature of the data considered. We analyzed the human interaction network using only eukaryotic genomes, whereas bacterial genomes are much more dynamic, readily losing and gaining genes (Hao and Golding 2004; and many others). In such genomes, the phylogenetic signal is readily lost for any one protein, and the coevolution signal is strengthened when lateral gene transfer events involving multiple genes take place. In the eukaryotic genomes we have considered, coevolution appears to be more stable and readily measurable from simply considering pairs of genes.

The MMM approach also allowed us to discover coevolution that was present in subtrees containing human, that did not necessarily extend over the entire phylogenetic diversity covered by the 20 genomes considered. In the vast majority of cases, the size of the best-matching submatrix was smaller than the smaller of the two matrices in the pair being considered (data not shown). This could be an indication that coevolution (and perhaps protein interaction) is not apparent from those lineages not included in the submatrix, even when both proteins are present in those lineages. Besides allowing us to identify more lineage-specific coevolution, the fact that MMM allows us to compare unequally sized matrices conveniently helps us to identify coevolution even when members of the gene family have been misidentified either owing to alignment errors or clustering errors.

In the present analysis of human protein coevolution networks, we did not make use of MMM's full potential. An especially promising avenue of research will be to tease out putative interaction partners from complex assemblages of paralogs. Gene duplication in a pair of gene families can potentially lead to the specialization of individual members from one family to their counterparts in the other family. Here we would expect the phylogenetic signal to be less important, overwhelmed instead by the intricate duplication histories of member genes. The principal coevolutionary signal may then represent the structural interadaptations to maintain interactions or to create new ones. Additionally, we could also start to study in more detail the coevolution of isoforms and variants, as we have discussed above for pyruvate kinase, that may have alternate roles and interaction partners. Since MMM returns not only the top-scoring largest-common submatrix (as used here) but all such common submatrices ranked by score, we could explore alternative coevolutionary relationships, and the possibility of multiple interaction partners, in a more in-depth analysis of select full-length and protein domain families.

We chose here to focus on networks of coevolving proteins that included a human representative. Since we also focused on the most strongly coevolving proteins, these represent interactions that have been maintained over long periods of evolutionary time in many species and are thus likely to be of fundamental importance to cellular function. We found that biochemical pathways

are maintained by coevolution but that these proteins are more likely to follow the overall phylogenetic signal. On the other hand, physical interactions show stronger coevolution and are more specific, and our ability to discover them increased when controlling for phylogeny, indicating that the rate of amino acid substitutions in a protein influences the rate of evolution of a few interacting partners. The coevolution networks presented here generate many hypotheses for potential functional protein interactions, contributing to models of pathways and interaction networks that will help us to understand the intricacies of living systems, and why such systems sometimes fail.

Methods

The MatrixMatchMaker algorithm

MatrixMatchMaker (MMM) implements a strategy to detect sequence coevolution from a pair of distance matrices (typically of different size) by finding their largest common submatrix. Conceptually, the submatrix returned by the program represents the largest common subtree. For each triplet of distances between sequences A, B, and C (AB, AC, BC) in matrix 1, compatible triplets of distances between sequences W, X, and Y (WX, WY, XY) are sought in matrix 2, where a triplet is compatible if AB/WX ≈ AC/WY ≈ BC/XY. Using distance ratios allows the two matrices to have different scales, although the scale (rate of evolution) within each matrix is assumed to be internally consistent.

The comparison of a pair of distance ratios such as AB/WX with AC/WY is achieved by requiring that a comparison statistic, $|(AB \times WY) - (WX \times AC)|/[(AB \times WY) + (WX \times AC)]$, be less than a user-supplied tolerance parameter ($-a$). Note that there may be more than one triplet from matrix 2 compatible with (AB, AC, BC), and each such case is pursued. For each such compatible triplet found in matrix 2, a recursive algorithm is then invoked that progressively adds a remaining sequence from matrix 1 with a compatible placement within matrix 2, if any. The compatibility of the new sequence is judged against every previously discovered pair of sequences within the growing match, using the triplet comparison scheme described above. If the set of correspondences (reflecting the best-matching subtree) resulting from such a recursion is at least as large as the largest set of correspondences found to date, it is retained. We thus seek the largest common submatrix satisfying the match criteria, at the expense of smaller though possibly better-matching submatrices.

At the end of the run, the set of matches is reported that represents the largest common submatrices given the user-supplied tolerance for distance ratio inequality. When this tolerance is set very low, fewer smaller and more precise submatrices will match, whereas when the tolerance is set high, more sequences will be encompassed within a more loosely matching framework. Several different sets of correspondences (submatrices) may be reported that meet the criteria described above, but some sets among these will represent better matches than others. Consequently, we score each set of correspondences by its root mean square deviation (RMSD) computed from all of its subsumed triplet distance comparison statistics. The maximum such deviation is a function of the supplied tolerance, since all reported matches must have fit within this tolerance in order to be reported. If desired, the user may specify a further constraint requiring matches between matrix 1 and matrix 2 to be from the same taxon, properly encoded within the taxon name. The user can also specify to consider only matches that contain a particular taxon.

The output of MMM is the maximum matching submatrix size, and for each possible solution of that size, its RMSD score and the list of paired-up protein names. The main score that we considered for this analysis was

$$\text{MMM score} = \text{Match size} + \text{RMSD score},$$

where

$$\text{RMSD score} = (1 - \text{RMSD})/(1 - \text{tolerance}).$$

Duplicate sequences are ignored by the algorithm, as distances of zero do not increase the score (as the absence of evolution in two different protein families cannot be taken as evidence for their coevolution). Additionally, to accommodate the nonlinear nature of evolutionary distances, the user-supplied tolerance can be modified as a function of the values in the matrix (the distances) in a stepwise manner. The tolerance is set to increase linearly from 0 to its maximum (set by the user as $-a$) in the distance interval [0, s1], is at its maximum in the second interval [s1, s2], and decreases toward 0 in [s2, s3]. The tolerance is set to zero in [s3, 1], which allows us to ignore very large distances for which the sequence alignments are unreliable.

Match potential represents the largest score that MMM could return, if thresholds were ignored. For reasons of efficiency (an MMM run without thresholds could be slow), we perform this computation in two steps. First, we compute the sum of the minimum pairwise counts for each taxon in common to the pair of matrices. This is possibly an overestimate since some pairings may be prohibited in a run of MMM if some of the distances in the matrices are zero, and as a consequence of the stepwise linear function. We therefore correct the match potential by subtracting the difference between a run without and with constraints on distance values, at the same tolerance threshold.

Removing the phylogenetic signal

Additional evidence for more-specific interactions would appear if the matrices were compatible with each other, but not with the overall phylogeny. The common phylogenetic signal could be measured in different ways. One approach would be to use a phylogenetic marker such as rRNA and obtain a matrix from the derived tree. Instead of using such a single phylogenetic marker, we calculated an average distance matrix for all the proteins considered. The MMM score or Pearson correlation of this matrix against all proteins would then indicate their propensity to follow the overall phylogenetic signal. However, since we necessarily exclude distances between paralogs within the same species in calculating such an average matrix, this method cannot be exact.

We also considered another approach, similar to what we used to remove phylogenetic signal when considering the coevolution of amino acids within a protein using Dependency (Tillier and Lui 2003). In this case, the score for a protein pair is reduced by the average of the scores against proteins with identical match potential. Although this approach would not be possible if we were only considering a pair of proteins in isolation, it is easily applicable in our all-by-all analysis of the HomoloGene clusters and has the added advantages of being able to consider paralogs and not to require an accurate tree phylogeny.

Strengthening network ties with MMMC

Valencia's group (Juan et al. 2008) found an increased accuracy of the mirror tree approach for the *E. coli* interaction network, by correlating the scores of protein pairs. This approach was thought to increase the coevolutionary signal by considering higher-order interactions rather than simply pairs of proteins, and we have

incorporated such a tactic in our MMM analysis of the human protein coevolution network. Our accessory program MMMC calculates the Pearson correlation coefficient between proteins based on the profile of their MMM scores, and derivatives thereof, with all the other proteins.

A listing of MMM output files is supplied to MMMC for processing, typically representing the results of each pairwise comparison of matrices in a collection. For convenience, we label each score from A through G:

A = MMM Score;
B = A/(1 + Match potential F);
C = A − Mean score within its class of Match potential F;
D = C/Match potential F;
E = Difference between Score A and max(second-best Score A for matrix *i*, second-best Score A for matrix *j*);
F = Match potential;
G = Match size − Size of match to average matrix (which is simply composed of the average intertaxon distance computed from the set of distance matrices under consideration).

Human proteins and homologs

To construct the human coevolution network, we used the HomoloGene database (Wheeler et al. 2008) as the source of sequence data. Although there are many other such databases that can be used for this purpose, we liked this one as it is constructed solely based on clustering by BLAST scores of the sequences from sequenced eukaryotic genomes (20, at the time of the analysis: *Anopheles gambiae*, *Arabidopsis thaliana*, *Ashbya gossypii*, *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Kluyveromyces lactis*, *Magnaporthe grisea*, *Mus musculus*, *Neurospora crassa*, *Oryza sativa*, *Pan troglodytes*, *Plasmodium falciparum*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*). The clusters thus obtained contain both orthologs and paralogs, and since a feature of MMM is its ability to consider paralogs, we wanted these included in our analysis. The distribution of distances between proteins in the database also seemed well suited for this analysis.

We downloaded 19,895 clusters containing at least one human sequence from HomoloGene, but we considered only the 3471 that contained more than 10 sequences. This was done to reduce the number of pairwise comparisons (from $\sim 300 \times 10^6$ to $\sim 6 \times 10^6$), but more importantly because we require some sequence diversity for any coevolution method. We obtained multiple sequence alignments for all clusters using MAFFT (Katoh and Toh 2008), because we had previously shown it to be one of the most accurate alignment programs (Nuin et al. 2006). Distance matrices were then obtained using ClustalW (Thompson et al. 1994) or Protdist (from PHYLIP) (Felsenstein 1989). We ran MMM with these 6×10^6 comparisons requiring at least one human protein pair as part of the solution. Only the top-scoring solution was considered from among those returned by MMM.

Data set of known interactions

A human protein interaction was considered as "known" if it was found in one or more of the following interaction databases: NCBI Gene (Wheeler et al. 2008), HPRD (Peri et al. 2003; Mishra et al. 2006), or BioGRID (Breitkreutz et al. 2008), or if any homologs in other species found in the same HomoloGene cluster as the human protein was found in those databases (interologs). For the proteins under consideration, with HomoloGene clusters containing at least one human protein and at least 11 sequences, we found 19,854 interactions. Additional annotation was obtained from the KEGG (Kanehisa et al. 2004) and GO (Ashburner et al. 2000) databases.

Acknowledgments

Funding for this work was provided by a grant from the Natural Sciences and Engineering Council of Canada to E.R.M.T. E.R.M.T. is the Canada Research Chair in Analytical Genomics.

References

Altenberg B, Greulich KO. 2004. Genes of glycolysis are ubiquitously overexpressed in 24 cancer classes. *Genomics* **84:** 1014–1020.
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29.
Bader GD, Hogue CW. 2002. Analyzing yeast protein–protein interaction data obtained from different sources. *Nat Biotechnol* **20:** 991–997.
Barabási A, Oltvai Z. 2004. Network biology: Understanding the cell's functional organization. *Nat Rev Genet* **5:** 101–113.
Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS, et al. 2009. An experimentally derived confidence score for binary protein–protein interactions. *Nat Methods* **6:** 91–97.
Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, et al. 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36:** D637–D640.
Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, et al. 2005. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433:** 531–537.
Christofk HR, Vander Heiden MG, Wu N, Asara JM, Cantley LC. 2008. Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* **452:** 181–186.
Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, Rual JF, Borick H, Braun P, Dreze M, et al. 2009. Literature-curated protein interaction datasets. *Nat Methods* **6:** 39–46.
Dekker C, Stirling PC, McCormack EA, Filmore H, Paul A, Brost RL, Costanzo M, Boone C, Leroux MR, Willison KR. 2008. The interaction network of the chaperonin CCT. *EMBO J* **27:** 1827–1839.
Felsenstein J. 1989. PHYLIP (Phylogeny Inference Package) version 3.2. *Cladistics* **5:** 164–166.
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* **296:** 750–752.
Fraser H, Wall D, Hirsh A. 2003. A simple dependence between protein evolution rate and the number of protein–protein interactions. *BMC Evol Biol* **3:** 11. doi: 10.1186/1471-2148-3-11.
Fryxell KJ. 1996. The coevolution of gene family trees. *Trends Genet* **12:** 364–369.
Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141–147.
Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736.
Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol* **299:** 283–293.
Hakes L, Lovell SC, Oliver SG, Robertson DL. 2007. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci* **104:** 7999–8004.
Hao W, Golding GB. 2004. Patterns of bacterial gene movement. *Mol Biol Evol* **21:** 1294–1307.
Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415:** 180–183.
Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y. 2000. Towards a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci* **97:** 1143–1147.
Izarzugaza JM, Juan D, Pons C, Pazos F, Valencia A. 2008. Enhancing the prediction of protein pairings between interacting families using

orthology information. *BMC Bioinformatics* **9:** 35. doi: 10.1186/1471-2105-9-35.
Jordan IK, Wolf YI, Koonin EV. 2003. No simple dependence between protein evolution rate and the number of protein–protein interactions: Only the most prolific interactors tend to evolve slowly. *BMC Evol Biol* **3:** 1. doi: 10.1186/1471-2148-3-1.
Jothi R, Przytycka TM. 2008. Computational approaches to predict protein–protein and domain–domain interactions. In *Bioinformatics algorithms: Techniques and application* (eds. I Mandoiu and A Zelikovsky), pp. i–xxxi. Wiley Book Series on Bioinformatics. Wiley, New York.
Jothi R, Kann MG, Przytycka TM. 2005. Predicting protein–protein interaction by searching evolutionary tree automorphism space. *Bioinformatics* **21:** i241–i250.
Jothi R, Cherukuri PF, Tasneem A, Przytycka TM. 2006. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *J Mol Biol* **362:** 861–875.
Juan D, Pazos F, Valencia A. 2008. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci* **105:** 934–939.
Jurica MS, Mesecar A, Heath PJ, Shi W, Nowak T, Stoddard BL. 1998. The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure* **6:** 195–210.
Kalaev M, Smoot M, Ideker T, Sharan R. 2008. NetworkBLAST: Comparative analysis of protein networks. *Bioinformatics* **24:** 594–596.
Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32:** D277–D280.
Kann MG, Jothi R, Cherukuri PF, Przytycka TM. 2007. Predicting protein domain interactions from coevolution of conserved regions. *Proteins* **67:** 811–820.
Kann MG, Shoemaker BA, Panchenko AR, Przytycka TM. 2009. Correlated evolution of interacting proteins: Looking behind the mirrortree. *J Mol Biol* **385:** 91–98.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9:** 286–298.
Kim WK, Bolser DM, Park JH. 2004. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics* **20:** 1138–1150.
Maere S, Heymans K, Kuiper M. 2005. BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **15:** 3448–3449.
Mika S, Rost B. 2006. Protein–protein interactions more conserved within species than across species. *PLoS Comput Biol* **2:** e79. doi: 10.1371/journal.pcbi.0020079.
Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, et al. 2006. Human protein reference database—2006 update. *Nucleic Acids Res* **34:** D411–D414.
Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y, Wang X. 1994. Co-evolution of ligand–receptor pairs. *Nature* **368:** 251–255.
Nathanson L, Deutscher MP. 2000. Active aminoacyl-tRNA synthetases are present in nuclei as a high molecular weight multienzyme complex. *J Biol Chem* **275:** 31559–31562.
Nuin PAS, Wang Z, Tillier ERM. 2006. The accuracy of several multiple sequence alignments for proteins. *BMC Bioinformatics* **7:** 471. doi: 10.1186/1471-2105-7-471.
Park SG, Ewalt KL, Kim S. 2005. Functional expansion of aminoacyl-tRNA synthetases and their interacting factors: New perspectives on housekeepers. *Trends Biochem Sci* **30:** 569–574.
Pazos F, Ranea JAG, Juan D, Sternberg MJE. 2005. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol* **352:** 1002–1015.
Pazos F, Juan D, Izarzugaza JM, Leon E, Valencia A. 2008. Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol* **484:** 523–535.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci* **96:** 4285–4288.
Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13:** 2363–2371.
Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schächter V, et al. 2001. The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409:** 211–215.
Ramani AK, Marcotte EM. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* **327:** 273–284.
Rives A, Galitski T. 2003. Modular organisation of cellular networks. *Proc Natl Acad Sci* **100:** 1128–1133.
Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. 2005. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437:** 1173–1178.
Sato T, Yamanishi Y, Kanehisa M, Toh H. 2005. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21:** 3482–3489.
Sato T, Yamanishi Y, Horimoto K, Kanehisa M, Toh H. 2006. Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. *Bioinformatics* **22:** 2488–2492.
Schwikowski B, Uetz P, Fields S. 2000. A network of protein–protein interactions in yeast. *Nat Biotechnol* **18:** 1257–1261.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504.
Shoemaker BA, Panchenko AR. 2007. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* **3:** e43. doi: 10.1371/journal.pcbi.0030043.
Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, Hirozane-Kishikawa T, Hao T, Sahalie JM, Venkatesan K, Gebreab F, et al. 2009. Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. *Nat Methods* **6:** 47–54.
Snel B, Bork P, Huynen M. 2002. The identification of functional modules from genomic association of genes. *Proc Natl Acad Sci* **99:** 5890–5895.
Souès S, Kann ML, Fouquet JP, Melki R. 2003. The cytosolic chaperon CCT associates to cytoplasmic microtubular structures during mammalian spermiogenesis and to heterchromatin in germline and somatic cells. *Exp Cell Res* **288:** 363–373.
Spirin V, Mirny L. 2003. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci* **100:** 12123–12128.
Sprinzak E, Sattath S, Margalit H. 2003. How reliable are experimental protein–protein interaction data? *J Mol Biol* **327:** 919–923.
Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al. 2005. A human protein–protein interaction network: A resource for annotating the proteome. *Cell* **122:** 957–968.
Steták A, Veress R, Ovádi J, Csermely P, Kéri G, Ullrich A. 2007. Nuclear translocation of the tumor marker pyruvate kinase M2 induces programmed cell death. *Cancer Res* **67:** 1602–1608.
Tabancay AP Jr, Forsburg SL. 2006. Eukaryotic DNA replication in a chromatin context. In *Current topics in developmental biology* (ed. GP Schatten), Vol. 76, pp. 129–184. Academic Press, San Diego.
Tanaka T, Harano Y, Sue F, Morimura H. 1967. Crystallization, characterization and metabolic regulation of two types of pyruvate kinase isolated from rat tissues. *J Biochem* **62:** 71–91.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.
Thulasiraman V, Yang CF, Frydman J. 1999. In vivo newly translated polypeptides are sequestered in a protected folding environment. *EMBO J* **18:** 85–95.
Tillier ERM, Lui TW. 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* **19:** 750–755.
Tillier ERM, Li G, Tillo D, Biro DL. 2006. Maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins* **63:** 822–831.
Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. 2000. A comprehensive analysis of protein–protein interactions of *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.
Valencia A, Pazos F. 2003. Prediction of protein–protein interactions from evolutionary information. *Methods Biochem Anal* **44:** 411–426.
van Kesteren RE, Tensen CP, Smit AB, van Minnen J, Kolakowski LF, Meyerhof W, Richter D, van Heerikhuizen H, Vreugdenhil E, Geraerts WP. 1996. Co-evolution of ligand–receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. *J Biol Chem* **271:** 3619–3626.
van Rijsbergen CV. 1979. *Information retrieval*, 2nd ed. Butterworth, London.
Veerassamy S, Smith A, Tillier ERM. 2003. A transition probability model for amino acid substitutions from BLOCKS. *J Comput Biol* **10:** 997–1010.
Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, et al. 2009. An empirical framework for binary interactome mapping. *Nat Methods* **6:** 83–90.
Vértessy BG, Bánkfalvi D, Kovács J, Löw P, Lehotzky A, Ovádi J. 1999. Pyruvate kinase as a microtubule destabilizing factor in vitro. *Biochem Biophys Res Commun* **254:** 430–435.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417:** 399–403.
Waddell PJ, Kishino H, Ota R. 2007. Phylogenetic methodology for detecting protein interactions. *Mol Biol Evol* **24:** 650–659.
Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287:** 116–122.
Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **36:** D13–D21.
Wuchty S, Almaas E. 2005. Peeling the yeast interaction network. *Proteomics* **5:** 444–449.
Wuchty S, Barabási AL, Ferdig MT. 2006. Stable evolutionary signal in a yeast protein interaction network. *BMC Evol Biol* **6:** 8. doi: 10.1186/1471-2148-6-8.
Yeang CH, Haussler D. 2007. Detecting coevolution in and among protein domains. *PLoS Comput Biol* **3:** e211. doi: 10.1371/journal.pcbi.0030211.
Yellaboina S, Dudekula DB, Ko MSh. 2008. Prediction of evolutionarily conserved interologs in *Mus musculus*. *BMC Genomics* **9:** 465. doi: 10.1186/1471-2164-9-465.
Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, et al. 2008. High-quality binary protein interaction map of the yeast interactome network. *Science* **322:** 104–110.

Received February 8, 2009; accepted in revised form June 9, 2009.

The human protein coevolution network

Elisabeth R.M. Tillier and Robert L. Charlebois

Genome Res. 2009 19: 1861-1871 originally published online August 20, 2009
Access the most recent version at doi:10.1101/gr.092452.109