Optimization criteria and biological process enrichment in homologous multiprotein modules

Luqman Hodgkinson and Richard M. Karp

Center for Computational Biology and Division of Computer Science, University of California, Berkeley, CA 94720

Contributed by Richard M. Karp, May 7, 2013 (sent for review January 3, 2013)

Biological process enrichment is a widely used metric for evaluating the quality of multiprotein modules. In this study, we examine possible optimization criteria for detecting homologous multiprotein modules and quantify their effects on biological process enrichment. We find that modularity, linear density, and module size are the most important criteria considered, complementary to each other, and that graph theoretical attributes account for 36% of the variance in biological process enrichment. Variations in protein interaction similarity within module pairs have only minor effects on biological process enrichment. As random modules increase in size, both biological process enrichment and modularity tend to improve, although modularity does not show this upward trend in modules with size at most 50 proteins. To adjust for these trends, we recommend a size correction based on random sampling of modules when using biological process enrichment or other attributes to evaluate module boundaries. Characteristics of homologous multiprotein modules optimized for each of the optimization criteria are examined.

Vestiges of similarity surviving evolution, homology, appear in distantly related taxa at many levels. Algorithms have been designed to detect homology at the level of multiprotein modules (1–5). These algorithms have been evaluated using overlap with literature-based protein complexes (4) and by graph theoretical measures (5); however, the most popular evaluation metric by far is Gene Ontology biological process enrichment.

Gene Ontology (6) stores a database of protein annotations that include participation in biological processes. Methods for biological process enrichment attempt to determine whether the module proteins have more similarity in biological process annotations than would be expected by chance. The most widely used method for biological process enrichment calculates the probability of obtaining at least as many proteins with the observed annotations if the module proteins were selected at random from a background set of proteins (7). Several variations on this basic idea have been introduced, for example, to consider the graphical structure of the ontology (8). Biological process enrichment algorithms have been implemented in Ontologizer (9).

The consolidated protein interaction databases contain a wealth of knowledge from numerous experiments, experimenters, and experimental assays. All assays have limitations on accuracy and precision, and the experimental protocols often differ between experiments (10). Searching for homologous modules in these data is a daunting challenge. Optimization criteria that may assist this search are of great interest.

In this study, we examine potential optimization criteria for detecting homology between Homo sapiens and Drosophila melanogaster at the level of multiprotein modules. We explore whether optimizing these attributes contributes significantly to biological process enrichment of the modules. After generating pairs of modules containing proteins homologous between the species, we consider different types of attributes: graph theoretical attributes intrinsic to a single module and homology attributes that measure similarity between homologous modules at lower levels. We consider the appropriateness of biological process enrichment for measuring the quality of module boundaries across modules of different sizes.

Modularity, defined as the fraction of interactions within the module among all interactions in which the module proteins participate, has been shown to lead to significant biological process enrichment (5). Graph conductance (11), a closely related concept, has been used to detect modularity in protein interaction networks not restricted to homologous modules (12). A similar definition of modularity (13) defined for a partition of the network into modules has been used with good results (14). These promising results cause modularity to be a candidate for predicting biological process enrichment.

Criticisms of modularity include claims that modules are not sufficiently dense (15). This prompts us to include a measure of density (i.e., how similar a module is to a clique) as a possible attribute for predicting biological process enrichment. Density is defined as the fraction of edges in the module over the number of edges if the module were a clique.

Two obvious attributes to consider are the number of interactions in the module and the module size. If these attributes are important, does biological process enrichment improve without bound as these quantities increase, or are there particular ranges that optimize biological process enrichment? Improvement without bound would indicate a limitation in the use of enrichment to evaluate module boundaries.

Modules vary in the diversity of proteins comprising them. It seems plausible a priori that modules with many paralogous proteins may exhibit more significant biological process enrichment than those with fewer paralogous proteins due to inherent similarity of function. However, it could also be possible that modules with diverse protein composition would have more significant biological process enrichment because each protein could perform a different molecular function in the biological processes associated with the module.

Protein interaction homology manifests itself through topological similarity in the protein interaction networks across species. Several attempts have been made to use protein interaction homology as a guiding principle to detect homologous modules, with the idea being that network regions containing homologous patterns of protein interactions are more likely to represent modules with important functions (3, 4). In Materials and Methods, we define a model of protein interaction homology and study its effects on biological process enrichment.

Results

Correlations and Regressions of Attributes. After generating 2,583 homologous module pairs at random as described in Materials and Methods, 2,583 modules per species, we computed six attributes for each module pair: modularity, density, number of interactions, module size, protein richness, and protein interaction similarity. These attributes are described in Materials and Methods. The number of interactions and module size were highly correlated with Pearson’s product-moment correlation coefficient, \( r^2 = 0.91 \); thus, we divided the number of interactions by module size, yielding an alternate measure of density that we call linear...
density following Melancon (16), and we did not consider the number of interactions directly. Protein richness was also highly correlated with module size with Pearson’s product-moment correlation coefficient \( r^2 = 0.84 \); thus, we divided protein richness by module size to create an attribute that we call size-corrected protein richness. We then performed curvilinear regression of biological process enrichment on each of the six attributes individually, using first- and second-order terms. The adjusted \( R^2 \) values are listed in Table 1. The correlation matrix of the attributes used in the regressions is given in Table S1. Multiple curvilinear regression of biological process enrichment on subsets of the six explanatory variables, including interactions between variables and second-order terms, yielded the adjusted \( R^2 \) values given in Table 1.

The protein interaction similarity score described in Materials and Methods ranged from 0 to 1 on the random modules, with a mean of 0.29 and SD of 0.15. The correlations with biological process enrichment were \( r = -0.05 \) (Pearson’s product-moment correlation coefficient) and \( \rho = -0.01 \) (Spearman’s rank correlation coefficient).

Relative Importances of Attributes. Four methods to determine the relative importance of attributes in curvilinear regressions are described in Materials and Methods. Fig. 1 shows the results of these methods applied to our regressions. “Lindeman, Merenda, and Gold (LMG),” “Pratt,” and “Last” listed modularity as the most important attribute. “First” listed linear density as the most important attribute, with modularity as the second most important.

Binary Partition Tree. Curvilinear regression models apply a polynomial function of the attributes to the entire sample space. An alternate nonparametric approach uses a binary partition tree to model the data with different predictors for different discrete ranges of attribute values. A binary recursive partitioning algorithm described in Materials and Methods generated the decision tree in Fig. S1, with modularity, linear density, and size being the most important attributes for predicting biological process enrichment.

Optimizations. As described in Materials and Methods, the algorithm was modified to select only proteins that improved the biological process enrichment of the modules. Due to the long running time of 8–20 h per homologous module pair, we generated only 6 homologous module pairs in this way, yielding 12 modules. For each optimization criterion other than biological process enrichment, we generated 100 homologous module pairs optimized for that attribute as described in Materials and Methods. To answer a question posed in the Introduction, we optimized for both high protein richness and low protein richness.

Fig. 2A compares biological process enrichment of optimized modules with random modules after correcting for size as described in Materials and Methods, Statistical Methods. The average sizes of modules resulting from the optimizations are listed in Table S2. All comparisons in this subsection refer to medians. Biological process enrichment in each optimized set of modules was greater than in the random control set (Wilcoxon rank sum test, \( P < 10^{-8} \) in all cases except for density, where \( P = 0.0017 \)). Biological process enrichment in the enrichment-optimized set was the highest, as expected (pairwise Wilcoxon rank sum test, \( P < 10^{-9} \) in all cases). Optimization for modularity and linear density led to greater enrichment than optimization for density (Wilcoxon rank sum test, \( P = 0.0015 \) and \( P < 10^{-9} \), respectively). Optimizing for linear density led to greater enrichment than optimizing for modularity (Wilcoxon rank sum test, \( P = 0.0015 \)).

Fig. 2B compares modularity in the optimized sets. The enrichment-optimized set had greater modularity than the random control set (Wilcoxon rank sum test, \( P < 10^{-9} \)). The modularity-optimized set had greater modularity than the enrichment-optimized set (Wilcoxon rank sum test, \( P = 0.007 \)).

Fig. 2C compares linear density and density in the optimized sets. That both linear density and density are represented by the same plots follows from the size correction described in Materials and Methods, Statistical Methods and the definitions of density and linear density. For modules of fixed size, the denominators of both density and linear density are constants; thus, each module is placed the same number of SDs above or below the mean for both density and linear density.

Fig. 2D compares size-corrected protein richness in the optimized sets. The enrichment-optimized set has lower protein richness than the random set and most other optimized sets. This is as expected from the negative correlation between enrichment and protein richness in the random set.

The greedy algorithm used in this study was remarkably effective at optimizing for protein interaction similarity. For every conserved module pair in the optimized set, the protein interaction similarity score was 1.0, indicating perfect topological agreement in each module, with the average number of protein homology groups per module being 14.13. Even so, as shown in Fig. 2A, these modules did not exhibit greater biological process enrichment than modules optimized for modularity or linear density, which, as shown in Fig. 2E, had significantly less protein interaction similarity.

The enrichment-optimized modules tend to have higher values of both modularity and linear density than the random modules, but the modularity-optimized modules do not have high linear density (Fig. 2C) and the linear density-optimized modules do not have high modularity (Fig. 2B). This is consistent with the regression results in the random modules that found modularity and linear density to be complementary predictors of biological process enrichment.

From Fig. 2A, it is clear that minimizing protein richness improves biological process enrichment. Fig. 2E shows that the set of conserved modules optimized for low protein richness indeed has very low protein richness relative to the random set. Many of the conserved modules optimized for low protein richness had only one or two protein homology groups, which yielded perfect protein interaction similarity agreement as indicated in Fig. 2E. Fig. 2D shows that optimization for every attribute except high protein richness led to lower protein richness than in the random modules.

Fig. 1. Relative importance of optimization criteria for predicting biological process enrichment. Four relative importance methods were used: LMG (A), Pratt (B), Last (C), and First (D). Methods were applied to 2,583 randomly generated homologous module pairs. The 95% confidence intervals were generated from 1,000 bootstrap replicates. dens, density; lind, linear density; mod, modularity; rich, size-corrected protein richness; sim, protein interaction similarity; size, module size.
Size Effects. Fig. 3A shows that biological process enrichment tends to improve as the number of proteins increases. To test at larger size ranges, 100 homologous module pairs were generated using the algorithm in Materials and Methods and choosing random maximum sizes between 5 and 2,000. The results are displayed in Fig. S2A, showing that this trend continues up to a module size of 2,000 proteins. As shown in Fig. 3B, there is no strong correlation between modularity and size in modules up to 50 proteins. The correlation between modularity and size is only \( \rho = 0.108 \) (Spearman’s rank correlation coefficient) and \( r = 0.015 \) (Pearson’s product-moment correlation coefficient), which is statistically significant for Spearman’s rank correlation coefficient \( (P < 10^{-15}) \) but not for Pearson’s product-moment correlation coefficient \( (P = 0.28) \). However, as shown in Fig. S2B, at larger size ranges, modularity is highly correlated with the number of proteins. Linear density tends to increase with the number of proteins at all size ranges (Fig. 3C and Fig. S2C). Density, normalized protein richness, and protein interaction similarity tend to decrease with module size (Fig. 3D–F and Fig. S2D–F).

Discussion

Importances of the Optimization Criteria. When biological process enrichment is considered a gold standard for module composition, modularity and linear density are the most important optimization criteria and are complementary. Optimizing for protein interaction similarity has only a modest additional effect. The small difference in the adjusted \( R^2 \) between the model using only graph theoretical attributes and the full model shows that the attributes based on homology (size-corrected protein richness and protein interaction similarity) do not provide significant additional information for predicting biological process enrichment. Three graph theoretical attributes, modularity, linear density, and module size, together explain 35.7% of the variance in biological process enrichment, whereas the full model explains only an additional 2%.

Table 1. Adjusted \( R^2 \) values for curvilinear regressions of biological process enrichment on attributes and subsets of attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Adjusted ( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear density (lind)</td>
<td>0.179</td>
</tr>
<tr>
<td>Modularity (mod)</td>
<td>0.175</td>
</tr>
<tr>
<td>Module size (size)</td>
<td>0.112</td>
</tr>
<tr>
<td>Density (dens)</td>
<td>0.062</td>
</tr>
<tr>
<td>Size-corrected protein richness (rich)</td>
<td>0.031</td>
</tr>
<tr>
<td>Protein interaction similarity (sim)</td>
<td>0.011</td>
</tr>
<tr>
<td>mod, lind, size, dens, rich, sim</td>
<td>0.377</td>
</tr>
<tr>
<td>mod, lind, size, dens, rich</td>
<td>0.372</td>
</tr>
<tr>
<td>mod, lind, size, dens, sim</td>
<td>0.370</td>
</tr>
<tr>
<td>mod, lind, size, dens</td>
<td>0.356</td>
</tr>
<tr>
<td>mod, lind, size</td>
<td>0.357</td>
</tr>
<tr>
<td>mod, lind</td>
<td>0.313</td>
</tr>
<tr>
<td>mod, size</td>
<td>0.291</td>
</tr>
<tr>
<td>mod, dens</td>
<td>0.242</td>
</tr>
<tr>
<td>sim, rich</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Fig. 2. Box plots of optimized conserved modules against a random control. Biological process enrichment (A), modularity (B), linear density and density (C), size-corrected protein richness (D), and protein interaction similarity (E) are shown. The control consists of 2,583 randomly generated conserved module pairs. Each optimized set, except for biological process enrichment, consists of 100 optimized conserved module pairs. The biological process enrichment-optimized set consists of 6 optimized conserved module pairs.
Levels of Homology. That there is extensive homology between proteins in distantly related species is undisputed. Much attention has been given to the search for homology at higher levels, such as collections of proteins associated with modules; these efforts have met with considerable success (1–5). However, at the intermediate level, searching for homology of protein interactions, many studies, including this one, have found extreme limitations. It requires integrating over many protein interactions in flexible ways to obtain any signal of module-level homology. Interaction level similarity is a poor predictor of module-level homology.

That protein interaction similarity does not significantly predict biological process enrichment poses interesting questions. Is it that there is so little detectable protein homology at the protein interaction level that it cannot be used reliably to detect a signal of module homology? Zinman et al. (17) found some evidence that protein interaction similarity was more significant within modules than between modules, but their results may have been due to the increased density within modules that results from module search algorithms. Our study shows that evidence of protein interaction similarity leads to modest gains at best if the modules are evaluated in terms of biological process enrichment.

Lack of detectable protein interaction homology may be due either to rewiring of protein interaction networks during evolution (18) or to artifacts in the protein interaction assays and interaction inference protocols. In the first case, lack of protein interaction homology is an inescapable feature of evolution. In the second case, it is a limitation of current technologies. A protein interaction network, as retrieved from public databases, is a mosaic of experimental assays with different data processing protocols and often subtle differences in what it means to interact (10).

Gandhi et al. (19) compared protein interaction data in four distantly related species, *H. sapiens*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *D. melanogaster*, finding little evidence of protein interaction similarity. Considering the lack of evident similarity among protein interactions across distantly related taxa, Beltrao and Serrano (18) estimated the rate of rewiring of protein interaction networks to be extremely high, with ~1,000 protein interactions changing in *H. sapiens* each million years.

Methods using higher and lower levels of homology to predict protein interaction homology have met with more success. Protein interactions were predicted with a 30% success rate using protein similarity (20) and with a 40–52% success rate (2) when predictions were restricted to interactions between proteins homologous to proteins in dense complexes.

Module Boundaries. The number of interactions within a module always increases as the module increases in size; thus, the number of interactions, by itself, is not a useful optimization criterion unless the desired module size is fixed. Modularity, linear density, density, and biological process enrichment seem to provide reasonable criteria for determining optimal module size and composition. As shown in Fig. 3 and Fig. S2, however, biological process enrichment and linear density tend to increase with size in random modules. Modularity tends to increase in very large random modules but is distributed around a constant mean in random modules up to 50 proteins. When modules are reasonably small relative to the interactomes, modularity provides the ability to compare modules across different sizes to choose the best size and composition. Density tends to decrease with size in random modules.

**Improved Null Model for Multiprotein Modules.** To establish a size-corrected null model against which to evaluate the significance of biological process enrichment and other evaluation measures, a set of modules (conserved or otherwise depending on the study) should be generated by a random growth process. The distributions of evaluation metrics for these random modules, separated by size, can then be used to evaluate the modules generated by module detection algorithms. This procedure was used to display the data in Fig. 2 from the optimization algorithms.

**Measures of Biological Meaning.** Even with the proposed null model for multiprotein modules, biological process enrichment may retain limitations as a measure of biological meaning. Studies have shown that biological process enrichment tends to be an optimistic measure, with many multiprotein modules, even those unreasonable by other measures, often deemed significant (5, 21). This problem is at least partially addressed by the improved null model presented in this study. Subtracting a baseline value averaged over randomly generated modules of the desired size corrects for much of this overestimation. An alternative measure of biological meaning is semantic density (21), which can also be normalized by module size. It would be interesting to see whether the general trends observed with biological process enrichment hold for semantic density and other measures of biological meaning.

**Implications of Protein Richness Optimizations.** The fact that lower protein richness improves biological process enrichment suggests that many biological processes are implemented by homologous interacting proteins rather than proteins with diverse origins. Further testing is necessary to determine whether this is a genuine biological discovery or an artifact of Gene Ontology enrichment and protein interaction data.

**Optimizing for Biological Process Enrichment Directly.** When all proteins in the interactomes are sufficiently annotated with biological processes, one way to detect conserved modules is by directly optimizing for biological process enrichment. The conserved modules returned by this algorithm were highly enriched; one, for example, contained 32 of the 34 proteins in the interactome that were annotated with “negative regulation of transcription from RNA Polymerase II promoter,” leading to an enrichment *P* value of $6 \times 10^{-54}$. However, there are proteins that are not annotated or insufficiently annotated; with which modules should these proteins be associated? Also, what about interactomes for which annotations are few? Furthermore, the running time to optimize for biological process enrichment may be prohibitive if desiring a full compendium of conserved modules across the interactomes, whereas optimizing for graph theoretical attributes is much faster.

**Modularity, Density, and Linear Density.** Modularity measures whether the proteins in the module have been associated with each other to a significant extent relative to all protein interaction studies conducted on the proteins. Proteins that are highly studied with interactions to many other proteins require more interactions within the module to associate them with the module to the same extent. Proteins that have few interactions, either because they have...
been poorly studied or because they genuinely have few interactions, contribute to high modularity if a significant number of their interactions are with other module proteins. Modularity incorporates data from protein interaction studies between the module proteins and proteins in other modules; these data are not used by measures of density.

There has been concern that modularity may not be comparable across modules with different numbers of proteins (22), and Fig. 2B demonstrates some justification for this concern. However, as shown in Fig. 3B, our results do not indicate any strong correlation between modularity and size in modules up to 50 proteins. The effects described by Peng (22) and demonstrated in Fig. 2B are only noticeable at larger size ranges when the number of proteins in the module is a significant fraction of the number of proteins in the interactome.

The density of random modules decreases as their size increases. If density were important for predicting biological process enrichment, we should expect the smaller, denser modules to be more enriched, but the opposite is observed. Fig. S1 and Table 1 show that modularity and linear density, together, are the best predictors of biological process enrichment. Fig. 2B shows that the linear-density-optimized set does not achieve high modularity, and Fig. 2C shows that the modularity-optimized set does not achieve high linear density.

**Protein Interaction Similarity.** Protein interaction similarity does not seem to provide a strong signal for predicting biological process enrichment. To verify the robustness of this result, it is reasonable to try other models for protein interaction similarity. We repeated the regressions using a model of protein interaction similarity (5) that considers an interaction between homology groups to be conserved if the interacting proteins in the module for one species are homologous to at least one pair of interacting proteins in the module for the other species. Using only this attribute in a curvilinear regression, with first- and second-order terms, the adjusted $R^2$ was 0.028. When used in the full curvilinear regression, the adjusted $R^2$ was 0.374. These results are similar to those in Table 1. Advantages of the model described in Materials and Methods include increased sensitivity to changes in protein interactions and robustness to false-positive interactions in the protein interaction data.

**Limitations and Avenues for Further Study.** The full model with six attributes explains 37.7% of the variance in biological process enrichment, with 35.7% of the variance explained by modularity, linear density, and module size alone. The remaining 62.3% of the variance is not explained by any of the attributes considered. Other attributes that explain additional variance in biological process enrichment may be discovered from two complementary modeling approaches: models of graph theoretical properties that arise from the physical construction of protein interaction networks and models of homology that consider how protein interaction networks change during evolution.

**Materials and Methods**

**Study System.** iRefIndex (23) is a consolidated database of protein interactions that can be accessed with the PSIQUIC Web services protocol (24). Release 9.0 (December 16, 2011) provides unified protein interaction data through a publicly accessible PSIQUIC Web service: data compiled from IntAct (25), BioGRID (26), the Database of Interacting Proteins (27), the Human Protein Reference Database (28), the Biomolecular Interaction Network Database (29), the Molecular Interaction Database (30), MIPS (31), InnateDB (32), and MatrixDB (33). The UniProt Knowledgebase (UniProtKB) is a publicly available database of protein sequences accessible via Web services (34). BLAST is software that detects the sequence similarity of protein sequences and can be used to infer homology of proteins across taxa (35). BLAST is released as a stand-alone application to compute pairwise similarities between two large sets of proteins very quickly (35). Gene Ontology is a database of annotations for protein functions, including biological processes, that can be used for comparing known functions of proteins within modules (6).

**Study Species.** *H. sapiens* and *D. melanogaster* were chosen as study species because their protein interaction networks have been the most extensively studied of all metazoans. *H. sapiens* and *D. melanogaster* are separated by over 500 million years of evolution, with their most recent common ancestor being an early member of the Bilateria clade (36). It is estimated that *D. melanogaster* has ~14,000 protein-coding genes, whereas *H. sapiens* has between 20,000 and 25,000 protein-coding genes (37). Sixty-seven percent of the proteins in their protein interaction networks have homologous proteins in the other species with a BLAST E-value $<10^{-5}$ when averaged in both directions.

**Optimization Criteria.** Let $G = (V, E)$ be an interactome. A multiple cut is a set of proteins $M \subseteq V$, such that $|M| < |V|$ and such that the proteins in $M$ form an induced connected subgraph of the interactome. Modularity is defined as $\mu(M) = \frac{E(M) - \langle E(M) \rangle}{\Delta}$, where $E(M)$ is the set of interactions with both interactants in $M$, and $\langle E(M) \rangle$ is the set of interactions spanning $M$ and $\bar{V} \setminus M$. Of the interactions involving module proteins, modularity is the fraction contained entirely within the module. Linear density is defined as $\Delta(M) = \frac{|E(M)|}{|M|^2}$, the number of interactions between proteins in the module divided by the module size. Density is defined as $\Delta(M) = \frac{|E(M)|}{|M|^2}$. Of the total possible number of edges that could be in a module of a given size, density is the fraction of edges actually contained in the module.

To model the extent of protein richness, we placed each pair of module proteins with a BLAST E-value less than or equal to $10^{-5}$ in the same protein homology group, including proteins from both species if both species have homologs in a module. A homology group in the module proteins and proteins in other modules; these data are relatively contained in the module.

To model protein interaction similarity, we tested how well the interactions between protein homology groups were conserved in *H. sapiens* and *D. melanogaster*. For either species, if protein homology group A had a protein in the species and protein homology group B had a protein in the species, and the module for the species had density $\Delta$, then if the interactions between A and B were thrown down at random, we would expect $ab\Delta$ interactions between A and B in that species. We examined the actual number for each species to see whether it was above or below expectation. If they were both below expectation, then both homology groups were conserved; otherwise, we called this an agreement; otherwise it was a disagreement. This model considers that interactions between protein homology groups that do not agree quantitatively in both species, after adjusting for variations in protein interaction experimentation rates, indicate evolutionary change between the species with regard to this set of interactions. If there are $n$ protein homology groups in the module, there are $\frac{n(n-1)}{2}$ pairs of protein homology groups and this same number of possible agreements. However, a large percentage of the agreements are between protein homology groups that have no interactions between them in either species. We thus ignored agreements of this type and measured only agreements with at least one interaction between the protein homology groups in at least one species, reporting the fraction of the total possible number of these agreements.

**Computational Methods.** We obtained protein interaction data from iRefIndex, Release 9.0, for *H. sapiens* and *D. melanogaster*, consisting of 69,651 interactions on 12,692 proteins for *H. sapiens* and 38,731 interactions on 9,796 proteins for *D. melanogaster*. We filtered out any interactions that were derived from computational rather than experimental sources. Protein sequences for all proteins involved in iRefIndex interactions were retrieved from UniProtKB on March 26, 2012. Pairwise BLAST was run between the *H. sapiens* and *D. melanogaster* proteins to determine protein homologies. We recorded BLAST E-values for each pair. All data were retrieved and processed using the EasyProt software architecture (38).

Beginning at each protein, in turn, we randomly generated two sets of proteins, one set per species. Let A and B be sets of proteins for *H. sapiens* and *D. melanogaster*, respectively, initially containing one protein each. For a set of proteins S, let $|S|$ be the set of proteins interacting with any protein in S. Let $H(S)$ be the set of proteins homologous to any protein in S in the other species. Each step of the growing process added a protein $p$ to A randomly selected from $\{p : (A) : H(p) \cap |B| \neq 0\}$. Next, a protein in $H(p) \cap |B|$ was chosen at random and added to B unless $H(p) \cap |B| \subseteq B$. The rules of A and B were then reversed. This procedure was repeated until both sets were either of a randomly chosen maximum size between 3 and 50 proteins or until it was impossible to grow the sets further. The collection of sets with at least 5 proteins per species was retained for statistical analysis.

When optimizing for attributes, we started from randomly chosen proteins and generated a set of conserved modules using the above algorithm, adding only proteins that improved the chosen attribute of the module, allowing up...
to 50 proteins per module. If the modules reached a size where it was impossible to improve the chosen attribute by adding a pair of homologous proteins, one protein per species, both interacting with module proteins, the modules were then returned at this size. If the modules were able to grow beyond 50 proteins, they were returned at 50 proteins. When optimizing for density, the homologous module pairs were grown at random up to 5 proteins per species before the optimization criteria were applied due to the difficulty in finding larger cliques. For each attribute other than biological process enrichment, we optimized 100 homologous module pairs. Due to the long running time, we optimized only 12 module pairs for each species, pairing the module pairs for biological process enrichment. We used the randomly generated set of 5,166 modules from 2,583 homologous module pairs as a control.

To measure the similarity of known biological processes across proteins within modules, Gene Ontology enrichment values were calculated for each module, separately for each species, using Ontologizer (9). We used the latest release of the full ontology (March 26, 2012 release) and unfiltered UniProt annotation data (March 6, 2012 release) from the Gene Ontology Web site. Ontologizer was applied with the Term-For-Term setting with Bonferroni correction, and each annotation received a P value. For each module, the minimum P value obtained by the enrichment calculation over all terms was used in the statistical analyses, including the regressions, after being transformed using the negative base-10 logarithm.

**Statistical Methods.** All statistical analyses were performed using R (39). Multiple curvilinear regressions were fitted to predict the negative logarithm of the biological process enrichment P value using six attributes: modularity, linear density, depth, module size, size-corrected protein richness, and protein interaction similarity. For protein interaction similarity, each module in the pair was assigned the measurement from the pair. Each regression was fitted using all first- and second-order terms, including interactions between variables. A binary recursive partitioning algorithm (40) was used to determine the most important attributes for prediction of biological process enrichment P values. Unless indicated otherwise, all statistical results given have a P value < 0.01.

For relative importance, we used the relaimpo R package with the four methods: LMG, Pratt, Last, and First (41). LMG assigns each variable the average increase in $R^2$ when it is added to a regression model containing a subset of other variables (42). Pratt measures the product of the regression coefficient and the zero-order correlation for each variable (43). Starting with the full model, Last signs the reduction in $R^2$ when removing a variable from the model as the relative importance of the variable maximizing this quantity; this maximizing variable is then removed from the model, and the algorithm recourses on the smaller model. First is similar to Last but adds the variables from an empty model rather than removing them from a complete model, adding the variables in order of relative improvement in $R^2$. Bootstrapping with 1,000 bootstrap replicates was used to generate 95% confidence intervals for each method. The 5,166 random conserved modules generated as described in Materials and Methods, Computational Methods were binned by number of module proteins. For each bin, we computed the mean and SD for each attribute. For each attribute, each optimized module generated as described in Materials and Methods, Computational Methods was mapped to the number of SDs above or below the mean in the random modules, allowing direct comparison across modules of different sizes. All tests comparing medians of attributes in sets of modules were conducted using a two-sided Wilcoxon rank sum test with continuity correction.

**Acknowledgments.** We thank Garvesh Raskutti for valuable discussions concerning the work presented here.

Supporting Information

Hodgkinson and Karp 10.1073/pnas.1308621110

Stratification by Size
Despite the strong correlation between the number of proteins and number of interactions in the random modules (Pearson’s product-moment correlation coefficient, \( r^2 = 0.91 \)), for modules of a given size, there is considerable variation in the number of interactions. For the modules of each size, we performed a linear regression of enrichment on the number of interactions. All 53 least-squares best-fit lines had a positive slope. This shows that for any fixed size, improving the density or linear density tends to improve biological process enrichment. The Pearson partial correlation between the number of interactions and biological process enrichment holding number of proteins fixed was 0.28 with a \( P \) value \(<10^{-96}\). However, biological process enrichment and density are negatively correlated (Pearson’s product-moment correlation coefficient, \( r = -0.20 \); Spearman’s rank correlation coefficient, \( \rho = -0.26 \)). It is easier to find smaller dense regions, but improving density at the cost of size tends to diminish biological process enrichment.

---

**Fig. S1.** Binary partition tree for negative base-10 logarithm of biological process enrichment in random modules. lind, linear density; mod, modularity; size, module size.
Table S1. Correlation matrix of attributes on random conserved modules

<table>
<thead>
<tr>
<th></th>
<th>Enrich</th>
<th>Mod</th>
<th>Lind</th>
<th>Size</th>
<th>Dens</th>
<th>Rich</th>
<th>Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrichment (Enrich)</td>
<td>1.00000000</td>
<td>0.39940960</td>
<td>0.41253070</td>
<td>0.33289206</td>
<td>-0.20491436</td>
<td>-0.16507973</td>
<td>-0.04770611</td>
</tr>
<tr>
<td>Modularity (Mod)</td>
<td>0.39940960</td>
<td>1.00000000</td>
<td>0.14834113</td>
<td>0.01513133</td>
<td>0.08709516</td>
<td>0.03372056</td>
<td>0.06667026</td>
</tr>
<tr>
<td>Linear density (Lind)</td>
<td>0.41253070</td>
<td>0.14834113</td>
<td>1.00000000</td>
<td>-0.82738012</td>
<td>0.13505171</td>
<td>-0.1209851</td>
<td>-0.17010913</td>
</tr>
<tr>
<td>Module size (Size)</td>
<td>0.33289206</td>
<td>0.01513133</td>
<td>-0.82738012</td>
<td>1.00000000</td>
<td>-0.19738641</td>
<td>-0.48665568</td>
<td></td>
</tr>
<tr>
<td>Density (Dens)</td>
<td>-0.20491436</td>
<td>0.08709516</td>
<td>-0.13505171</td>
<td>-0.19738641</td>
<td>1.00000000</td>
<td>0.22562763</td>
<td>0.50260948</td>
</tr>
<tr>
<td>Size-corrected protein richness (Rich)</td>
<td>-0.16507973</td>
<td>0.03372056</td>
<td>-0.1209851</td>
<td>-0.48665568</td>
<td>0.22562763</td>
<td>1.00000000</td>
<td>-0.20942665</td>
</tr>
<tr>
<td>Protein interaction similarity (Sim)</td>
<td>-0.04770611</td>
<td>0.06667026</td>
<td>-0.17010913</td>
<td>-0.20942665</td>
<td>1.00000000</td>
<td>0.50260948</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Table S2. Average sizes of optimized modules

<table>
<thead>
<tr>
<th></th>
<th>Number of Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size-corrected protein richness maximized</td>
<td>48.11</td>
</tr>
<tr>
<td>Protein interaction similarity optimized</td>
<td>41.65</td>
</tr>
<tr>
<td>Linear density optimized</td>
<td>38.29</td>
</tr>
<tr>
<td>Enrichment optimized</td>
<td>36.58</td>
</tr>
<tr>
<td>Size-corrected protein richness minimized</td>
<td>29.83</td>
</tr>
<tr>
<td>Modularity optimized</td>
<td>28.54</td>
</tr>
<tr>
<td>No optimization</td>
<td>27.79</td>
</tr>
<tr>
<td>Density optimized</td>
<td>6.08</td>
</tr>
</tbody>
</table>