Supporting Materials and Methods: The PATHBLAST Algorithm

PATHBLAST Overview. PATHBLAST implements a scoring function and search algorithm to find high-probability pathway alignments between two protein interaction networks N_1 and N_2. We define a pathway alignment to consist of two paths, one from each network, in which proteins in the first path <A, B, C, D,…> pair with putative homologs occurring in the same order in the second path <a, b, c, d,…> (Fig. 1*a*; see main text). A particular homologous protein pair may not occur more than once per pathway alignment. The pathway alignment may include nonhomologous proteins by introducing "gaps" and "mismatches." A gap occurs when a protein interaction in one path skips over a protein in the other, whereas a mismatch occurs when two proteins at the same position in the alignment do not share sequence homology. Neither gaps nor mismatches may occur consecutively.

For scoring and search, it is convenient to combine the two protein interaction networks into a global alignment graph G (Fig. 1*b*; see main text). Each vertex in G represents a protein pair A/a between N_1 and N_2. Vertices A/a and B/b are connected by an undirected edge (of type direct, gap, or mismatch) if:

- Protein–protein interactions (A,B) and (a,b) are present in N_1 and N_2 [direct];
- (A,B) is present in N_1, and the distance between a,b in N_2 is 2 [gap in N_1];
- (a,b) is present in N_2, and the distance between A,B in N_1 is 2 [gap in N_2];
- A,B and a,b are connected at distance 2 in both N_1 and N_2 [mismatch]

A pathway alignment corresponds to a simple path P through G.

Scoring Function. We formulate a log probability score $S(P)$ that decomposes over the vertices v and edges e of path P,

$$S(P) = \sum_{v \in P} \log_{10} \frac{p(v)}{p_{random}} + \sum_{e \in P} \log_{10} \frac{q(e)}{q_{random}} \qquad (1)$$

where $p(v)$ is the probability of true homology within the protein pair represented by v, $q(e)$ is the probability that the protein–protein interactions represented by e are real, i.e., not false-positive errors. The background probabilities p_{random} and q_{random} are the expected values of $p(v)$ and $q(e)$ over all vertices and edges in G.

The value of $p(v)$ is computed using Bayes' rule, given the pairwise protein sequence similarity for the proteins in v expressed as a BLAST E value E_v,

$$p(v) = p(H \mid E_v) = \frac{p(E_v \mid H)\, p(H)}{p(E_v)} \qquad (2)$$

where H represents the event of true homology between the proteins represented by v. The probability distribution for $p(E_v)$ is taken as the frequency of each E value over all v in G (i.e., over all protein pairs; see Fig. 5). The probability distribution for $p(E_v|H)$ is based on E values within the subset of vertices for which both proteins are in the same

cluster of orthologous groups (COG) (1), a commonly accepted classification of true protein orthology.† The constant prior probability $p(H)$ is computed as the overall frequency of vertices with proteins that are in the same COG. Probability distributions are smoothed by using a monotone regression function (we used the pool-adjacent-violators algorithm as described in ref. 2) and indexed with the appropriate value of E_v.

The probability $q(e)$ of each edge is computed from the underlying probabilities of the protein–protein interactions it represents. By construction of the global alignment graph, if e is direct it represents two interactions (one from each network), whereas if e is a gap or mismatch it represents three or four interactions, respectively. Using recently published guidelines on the accuracy of protein interaction data (3), we roughly estimate the probability of each interaction i by the number of independent experimental studies reporting it and then compute $q(e)$ as the product of these probabilities. Introducing gaps and mismatches penalizes $q(e)$ because more interaction probabilities must be included in the product:

$$q(e) = \prod_{i \in e} \Pr(i) \qquad (3)$$

Number of studies	Pr(*i*)	# yeast interactions
1	0.1	9966
2	0.3	1597
≥ 3	0.9	1591

More complex interaction scoring functions are also possible such as a function that accounts not only for the number of experimental replicates but also for the type of experiment, e.g. whether the interaction was generated by using a two-hybrid or coimmunoprecipitation assay. However, rigorous evaluations of interaction data quality are still ongoing, and error models for protein–protein interaction data have yet to be developed. In this study we opted for a probability function based on relatively straightforward assumptions, although we note that $q(e)$ can become arbitrarily complex as information improves.

Alignment Procedure. We wish to identify the highest-scoring pathway alignment P^* of fixed length L (L vertices and $L - 1$ edges). If G is directed and acyclic, this can be accomplished in linear time (in the number of edges) by using dynamic programming, in which the highest-scoring path of length $l = 2 \ldots L$ ending in vertex v will have score

$$S(v,l) = \arg\max_{u \in parents(v)} \left[S(u,l-1) + \log \frac{p(v)}{p_{random}} + \log \frac{q(e_{u \to v})}{q_{random}} \right] \qquad (4)$$

and the base case is

$$S(v,1) = \log \frac{p(v)}{p_{random}} \qquad (5)$$

Because G is not generally acyclic, we first construct a sufficient number ($5L!$) of directed acyclic subgraphs and then use the dynamic programming method to compute the highest-scoring paths for each. To construct a particular acyclic subgraph G', we induce a random ordering on the vertices of G and remove all edges with sources that are lower in rank than their targets. On average, the highest-scoring path in G will be

preserved in G' once for every 2/L! subgraphs examined. For example, an expected 1/60th of the random subgraphs (because G is undirected) will contain the highest-scoring path of length L = 5. Fig. 6 charts the probability of finding the optimal path for different lengths L vs. the number of subgraphs examined.

As a necessary property, when this algorithm is used to identify pathway alignments between two identical protein networks (in both topology and protein sequence composition) the proteins aligned at each position of P* will be identical.

[†]Note from Fig. 5 that there is a positive but imperfect association between E values and orthology as defined by the COG database. Therefore, we have not used the BLAST E value as a direct measure of orthology but have made indirect use of E values by comparing their distributions among orthologous and nonorthologous proteins.

1. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000) *Nucleic Acids Res.* **28**, 33–36.
2. Haerdle, W. (1992) *Applied Nonparametric Regression* (Cambridge Univ. Press, Cambridge, U.K.).
3. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) *Nature* **417,** 399–403