Supporting Text

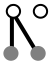
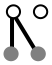
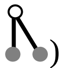
Model Testing. We tested the algorithm on simulated networks generated with a range of values for parameters P_+, P_-, P_i and P_{si}. For each parameter set, we initially create an ancestral network of N nodes. At each of the N possible self-interaction sites, we create a link with probability P_{si}. Between each of the ½N(N–1) pairs of proteins, we create an interaction with probability P_i. We simulate the WGD by duplicating the network in its entirety. If two proteins interacted with each other prior to the duplication, then all four pairs of their duplicates interact with each other in the duplicated network. If a protein was self-interacting, then its duplicates interact with each other and with themselves in the duplicated network.

We then simulate the divergence period. Every interaction that is present in the network is removed with probability P_-, and every interaction that is absent is created with probability P_+. The final network is a function of P_i and P_{si} (which determine the architecture of the pre-duplication graph) and P_+ and P_- (which describe the period of divergence).

We test the algorithm by checking whether it is able to determine the four parameters used in the construction when given a simulated network as its input. The algorithm successfully reconstructed P_i, P_{si}, P_+, and P_- for a wide range of parameter values. An example of the algorithm's performance on a simulated network is illustrated in SI Table 2.

Error Estimation. We simulate networks using the *S. cerevisiae* best-fit parameters and the method described above. We then use the fitting algorithm described in the text to extract those parameters from the simulated networks. Because of the finite network size, the best fit values of P_i, P_{si}, P_+, and P_- in each network realization is somewhat different than the input values. We estimate the uncertainty in our *S. cerevisiae* fit parameters to be the standard deviation associated with the simulated network fits.

Outlying Motifs. While in general the model is quite good in fitting the various motifs, there are some that lie outside the expected range. Particularly in the case of rare motifs, there is the possibility that the true frequency of the motif is masked by noise inherent in the proteomic data. Given enough statistical power, outlying motifs can suggest actual differences between the evolutionary process and our simple model. Such deviation from our model can offer interesting insight into the evolutionary process. Where the motifs are more frequent then expected it is possible that the motif is functional, and selectively preserved in the proteome (1). Similarly in the case of under-represented motifs, if the structure is for some reason unfavorable for the organism, those motifs will tend to disappear more rapidly than expected.

As an illustration, we point out the underrepresented motif,  . We denote the white nodes as A_1 and A_2, and the grey as B_1 and B_2. While it is possible that the present interactions are de novo, it is far more likely that they descend from an ancestral interaction between the parental proteins, A and B. Based on our fit, we expect about 24 of these motifs, yet we only observe 5. Note that in this particular motif, protein A_1 has retained interactions with both daughters of the other pair (B_1 and B_2), and presumably maintains the ancestral function with respect to the B pair. A_2 however, has lost its ancestral functionality with respect to B, and its post-duplication functionality with respect to B_1 and B_2. A possible explanation could be that these motifs tend to delete the protein that has lost its edges, and hence do not survive as ohnolog pair motifs (i.e.  → ).

Dose Dependent Model. We suggest a simple dose-dependent model consistent with the possibility that duplicated self-interacting proteins are selectively preserved (2). In the case of a self-interacting protein **A** duplicating into $\mathbf{A_1}$ and $\mathbf{A_2}$, there will be three possible protein complexes: $\mathbf{A_1A_1}$, $\mathbf{A_2A_2}$ and $\mathbf{A_1A_2}$ in a ratio of 1:1:2. If either gene develops a deleterious mutation that renders the complex non-functional, the total number of functional complexes will be reduced by a factor of 4, leaving only half the number present prior to duplication (SI Fig. 5). This mechanism will exert selective pressure on both ohnologs to keep them from acquiring deleterious mutations (3). This pressure could allow one of the ohnologs to acquire a mutation favoring the heterodimer, in which case the self-interacting nature of the proteins may eventually be lost.

This generation of paralogous interacting genes via the duplication of self-interacting proteins can contribute to the evolutionary formation of protein complexes (4). For example, the ohnologs PIP2 and OAF1 are transcription factors of the Zn_2Cys_6 zinc finger family of proteins. They are a WGD pair that interact with each other in the modern proteome, but do not interact with themselves (5). They presumably descend from an ancestrally self-interacting protein, from the WGD with a paralogous interaction, and subsequently lost their self-interactions. In broader contexts, hemoglobin α and β make up a duplicate pair. The α and β proteins interact with each other and with themselves to form the quaternary hemoglobin molecule (2 α and 2 β). Photosystem I could be yet another example of this phenomenon (6). The history of the hemoglobin complex and of photosystem I suggest that the preferential maintenance of ancestrally self-interacting duplicates may take place in other species as well (7).

Expanded Model. We presented three possible mechanisms to explain the discrepancy between the estimated ancestral value of P_{si}, and the modern one: first, that the ancestral network had a higher P_{si} than today's network (mechanism 1), second, the ancestral P_{si} was roughly the same as the modern one, but the probabilities of adding or deleting interactions between ohnologs differ from the background rate of adding and deleting interactions (mechanism 2), and third, that the ancestrally self-interacting proteins were selectively preserved in duplicate to the modern day (mechanism 3) (8). The model could be naturally extended to account for these different possible mechanisms.

Mechanism 2 depends upon a difference existing between the probability of adding/deleting an edge between ohnologs ($P_{+/-,ohnolog}$), and the probability of adding an edge between non-ohnologous proteins ($P_{+/-,non-ohnologs}$). Thus, by replacing $P_{+/-}$ with $P_{+/-,ohnolog}$, and $P_{+/-,non-ohnolog,}$, we can examine these differences.

Mechanism 1 and Mechanism 3 can be resolved by introducing a parameter distinguishing the retention probabilities of pairs descending from ancestrally self-interacting proteins (R_{si}) and from ancestrally non-self-interacting proteins ($R_{\neg si}$). When these retention probabilities are added, the P_{si} value that remains in the model represents the true self-interaction probability of the ancestral network, as opposed to the self-interaction probability of the genes whose duplicates have survived.

We find that while the extended model cannot fully resolve the values of the added parameters, it provides analytical relationships between them. The complete system with the added parameters completely decouples into two separate subsystems, one describing the interactions within pairs of ohnologs (‘ohnolog-interactions’), and one describing all other interactions (‘ordinary-interactions’). This is because the equations describing the ‘ohnolog-interactions’ no longer depend on P_i, $P_{+,non\text{-}ohnolog}$, or $P_{-,non\text{-}ohnolog}$. Thus all of the information the network contains about ‘ordinary interactions’ no longer contributes to the solution of the ‘ohnolog-interaction’ equations and vise versa. The equations describing the ‘ohnolog-interactions’ are therefore underdetermined. This subsystem is:

$$R_{si}P_{ancestral,si} + R_{\neg si}(1 - P_{ancestral,si}) = P_{dup} \qquad [2]$$

$$\frac{R_{si}P_{ancestral,si}(1 - P_{-,ohno\log}) + R_{\neg si}(1 - P_{ancestral,si})(P_{+,ohno\log})}{R_{si}P_{ancestral,si} + R_{\neg si}(1 - P_{ancestral,si})} = P_{ohno} \qquad [3]$$

where P_{ohno} is the observed probability of ohnolog interaction in the modern network, and P_{dup} is the observed probability of a gene being a member of an ohnologous pair.

Eq. **2** sums the frequencies of each kind of ancestral protein (self-interacting and non-self-interacting) multiplied by the retention probabilities of each type. The result is the observable frequency of WGD pairs in the modern network (P_{dup}). Eq. **3** sums the probabilities of interacting ohnologs resulting from the two types of ancestral proteins. This sum, normalized by the frequency of duplicates in the network, yields the observable frequency of interacting ohnologs (P_{ohno}). Note that the only relevant data to P_{ohno} are the probabilities of various temporal paths to interacting ohnologs.

The resulting decoupled subsystem describing the ‘ohnolog-interactions’ has 2 equations and 5 unknowns, and cannot be solved uniquely. Additional types of data may be used in the future together with these expressions to solve for these new parameter values, thereby distinguishing between the three possible mechanistic explanations.

SI References.

1. Milo, R, Shen-Orr, S, Itzkovitz, S, Kashtan, N, Chkovskii, D, Alon, U (2002) *Science* 298:824-827.

2. Hughes, T, Ekman, D, Ardawatia, H, Elofsson, A, Liberles, DA (2007) *Genome Biol* 8, 8:213.211 - 218:213.214.

3. Aury, J-M, Jaillon, O, Duret, L, Noel, B, Jubin, C, Porcel, BM, Segurens, B, Daubin, V, Anthouard, V, Aiach, N, *et al* (2006) *Nature* 444:171-178.

4. Pereira-Leal, JB, Levy, ED, Kamp, C, Teichmann, SA (2007) *Genome Biol* 8:R51.51-R51.12.

5. Wolfe, K (2001) *Nat Rev Genet* 2:333-341.

6. Ben-Shem, A, Frolow, F, Nelson, N (2004) *FEBS Letters* 564: 274-280.

7. Czelusniak, J, Goodman, M, Hewett-Emmett, D, Weiss ML, Venta, PJ, Tashian, RE (1982) *Nature* 298:297-300.

8. Wagner, A (2003) *Proc R Soc Lond B* 270:457-466.