

Cross-species analysis of biological networks by Bayesian alignment

Johannes Berg[†] and Michael Lässig

Institut für Theoretische Physik, Universität zu Köln, Zùlpicherstrasse 77, 50937 Cologne, Germany

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved May 22, 2006 (received for review March 27, 2006)

Complex interactions between genes or proteins contribute a substantial part to phenotypic evolution. Here we develop an evolutionarily grounded method for the cross-species analysis of interaction networks by alignment, which maps bona fide functional relationships between genes in different organisms. Network alignment is based on a scoring function measuring mutual similarities between networks, taking into account their interaction patterns as well as sequence similarities between their nodes. High-scoring alignments and optimal alignment parameters are inferred by a systematic Bayesian analysis. We apply this method to analyze the evolution of coexpression networks between humans and mice. We find evidence for significant conservation of gene expression clusters and give network-based predictions of gene function. We discuss examples where cross-species functional relationships between genes do not concur with sequence similarity.

cross-species network analysis | network statistics

Besides a wealth of genomic sequence information, molecular biology is accumulating more and more data probing the interactions between genes or proteins. Examples are regulatory interactions, where the expression level of one gene influences the expression of another gene, or interactions between proteins, where pairs of proteins bind to form dimers or multimers. Interactions between genes or their products are crucial for our understanding of biological functions. With the advent of experimental high-throughput methods, large-scale datasets of different organisms are becoming available, which can be analyzed by systematic cross-species comparison.

This article is devoted to developing an evolutionary rationale for biological network analysis. Because the interactions between genes are encoded in their genomic sequences, this may seem a rather straightforward generalization of established concepts in sequence analysis: evolution acts as a divergent process on the constituents of the network, which gradually reduces cross-species correlations of the network structure. Detecting these correlations requires an alignment procedure that can map functional units as network structures conserved by evolution as well as estimate the degree of divergence between species.

However, interaction networks evolve in a more heterogeneous and a more correlated way than sequences, which makes their cross-species comparison a considerably more challenging task. The interactions between proteins, for example, depend on the properties of a specific functional binding domain, which may evolve in a different way than the remainder of the protein sequence, with correlations to its binding domain in a different protein. Regulatory interactions can change by the evolution of regulatory DNA, which is expected to be different from that of coding DNA. Moreover, many sequence changes in a gene may be irrelevant for its interactions measured in a network. These different modes lead us to treat the evolution of the interactions within a network, i.e., its *link dynamics*, and the overall sequence evolution of its constituents, the *node dynamics*, as two independent modes of evolution. We describe these modes by simplified stochastic models and infer their relative contribution to network evolution by cross-species comparison. This dynamics

is also quite heterogeneous across the network, and we use the models of link and node dynamics to quantify the evolutionary conservation of putatively functional network modules.

Our evolutionary analysis is based on the *alignment* of networks, i.e., a mapping between their nodes, which also induces a mapping of their links. In the *Theory* part of this article, we develop a statistical theory of alignment for biological networks. We introduce a scoring designed to detect local functional correlations, which uses the similarities both of mapped link pairs and of node pairs. This scoring derives from the underlying link and node dynamics.

Various alignment and scoring procedures for biological networks have been discussed in recent articles. One type of method restricts the alignment to mutually homologous nodes, i.e., gene pairs with significant sequence similarity in different species. In this way, clusters of conserved interactions have been found in gene coexpression networks (1, 2) and in protein interaction networks (3, 4). A complementary approach is to align networks only by their link overlap, independently of node homology. Network motifs (5, 6) defined by families of mutually similar subgraphs in a larger network have been identified in this way (7) as well as the similarities between regulatory networks of different phages (8). These methods have been combined with their relative weights fixed *ad hoc* in ref. 9. A third method, called Pathblast (10, 11), evaluates the link similarity between networks along paths of connected nodes, using sequence alignment algorithms. It has been applied to cross-species comparisons of protein interaction networks (10). Similarly, the flux along the shortest paths in regulatory networks has been compared across species (8). Metabolic networks with few cycles have been analyzed by subtree comparison (12).

From an evolutionary point of view, these methods are heuristics containing different assumptions on the underlying link and node dynamics. Homology-based alignments are appropriate if the sequence divergence between the species compared is sufficiently small so that all pairs of functionally related nodes can be mapped by sequence homology. However, genes with entirely unrelated sequence may take on a similar function in different organisms, and hence have a similar position in the two networks. (Such so-called nonorthologous gene displacements are well known in metabolic networks (13–15).) On the other hand, alignments by link similarity alone altogether ignore the evolutionary information of the node sequences. Path-based alignment algorithms are well suited to networks with predominantly linear biological pathways such as signal-transduction chains. In other situations, however, it may be difficult to link the scoring parameters to evolutionary rates of link and node changes.

The alignment method presented in this paper is grounded on statistical models for the evolution of links and nodes. Alignments are constructed from link and node similarity treated on an equal footing. The relative weight of these score contributions is deter-

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

[†]To whom correspondence should be addressed. E-mail: berg@thp.uni-koeln.de.

© 2006 by The National Academy of Sciences of the USA

mined systematically by a Bayesian parameter inference. Nodes without significant sequence similarity are aligned if their link patterns are sufficiently similar. Conversely, nodes are not aligned despite their sequence similarity if their links, and hence their putative functional role, show a strong divergence between the two networks. Our method is rather general and can be applied both to networks with binary link strengths (as in the current large-throughput data for protein interactions) and to networks with continuous link strength (such as the coexpression data used in this study).

As an algorithmic problem, network alignment is clearly more challenging than sequence alignment, which can be solved by dynamic programming (16, 17). Already simpler problems such as matching two graphs by determining the largest common subgraph are NP-hard (18), which implies there is probably no polynomial-time algorithm. We have developed an efficient heuristic, by which network alignment is mapped onto a generalized quadratic assignment problem, which in turn can be solved by iteration of a linear problem (19).

In the second part of the article, we present a cross-species comparison of coexpression networks of *Homo sapiens* and *Mus musculus* as an example application of our method. In this type of network, the link between a pair of genes is given by the correlation coefficient of their expression profiles measured on an RNA microarray chip. We show that correlation networks are well suited for cross-species comparison: they are robust datasets even if individual expression levels cannot be compared with each other because the experimental conditions differ between species. The evolution of these networks results from the evolution of regulatory interactions between genes and from loss and gain of genes. High-scoring alignments between expression networks in human and mouse provide a quantitative measure of divergence between the two species. We find conserved network structures, related to clusters of coexpressed genes; similar findings are reported in refs. 1 and 4. However, the alignment found here differs from mere sequence homology. This finding leads to network-based predictions of gene functions, including functional innovations such as nonorthologous gene displacements.

Theory

Graphs and Graph Alignments. A graph A is a set of nodes with links between pairs of nodes. The graphs considered here are labeled by gene name, which is denoted by the node index $i = 1, \dots, N_A$, and are thus uniquely represented by the adjacency matrix $\mathbf{a} = (a_{ij})$. A graph is called *binary* if links are either absent ($a_{ij} = 0$) or present ($a_{ij} = 1$) and *weighted* if the link strengths a_{ij} take continuous values. The special case of a symmetric adjacency matrix is used to describe *undirected* graphs. For example, current high-throughput datasets of protein interactions, which do not specify the interaction strength, produce binary undirected graphs. Gene expression networks, whose links denote the mutual correlation coefficient between expression patterns of two genes, are weighted graphs with $-1 \leq a_{ij} \leq 1$.

A *local alignment* between two graphs A and B is defined as a mapping π between two subgraphs $\hat{A} \subset A$ and $\hat{B} = \pi(\hat{A}) \subset B$ as shown in Fig. 1a. The alignments of networks discussed in this paper are designed to display cross-species functional relationships between aligned node pairs. Because of gain or loss of genes in either species, not every gene in one network has a functional equivalent in the other, and the alignment algorithm has to determine the aligned subnetworks \hat{A} and \hat{B} with significant correlations. For the sake of algorithmic simplicity, we will discuss here only one-to-one mappings π , which is appropriate for most gene pairs but neglects multivalued functional relationships induced by gene duplications.

Link Dynamics and Link Score. An important statistical characteristic of a network is the link distribution $p^\ell(a)$, giving the probability that the link between a randomly chosen pair of nodes takes on value a .

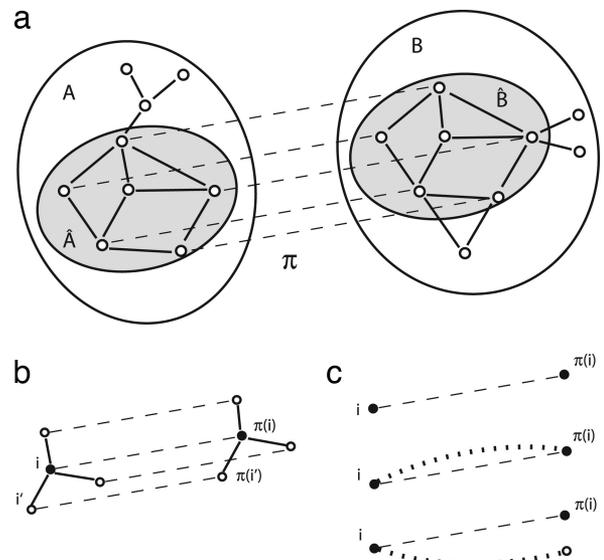


Fig. 1. Network alignments measure link and node similarity. (a) A local alignment π between two networks A, B is a one-to-one mapping (indicated by dashed lines) between nodes of the subsets \hat{A}, \hat{B} . (b) The local link score $S_{i,\pi(i)}^l$ evaluates all pairwise similarities between links a_{ij} and $b_{\pi(i)\pi(j')}$ (solid lines) for a given pair of aligned nodes. (c) The local node similarity score $S_{i,\pi(i)}^n$ evaluates the overlap of the alignment with the node similarity $\theta_{i,\pi(i)}$ (dotted line). Top to bottom: Aligned node pairs without similarity to any other node, with mutual node similarity, and with (at least one) node similarity mismatch.

The evolution of the link distribution can be modeled by a simple stochastic process, from which our link similarity scoring of an alignment is derived. In a binary network, the simplest form of link dynamics is a Markov process, which is fully determined by the rates of formation and loss of single links. Generalizing this dynamic to continuous links leads to a diffusion equation of the form

$$\partial_t p^\ell(a) = [\partial_a^2 g(a) - \partial_a f(a)] p^\ell(a). \quad [1]$$

The two terms on the right hand side describe the stochastic turnover and the average relaxation of links with coefficient functions $g(a)$ and $f(a)$, respectively. For mutual expression correlations between two genes in a microarray, this form can be derived from a stochastic model for loss and gain of regulatory interactions, each of which affects a random subset of the experiments, respectively, cell types. The cross-species correlations in pairs of evolutionarily related links a, b are contained in the joint distribution $q^\ell(a, b)$, which we write in the form

$$q^\ell(a, b) = p_A^\ell(a) p_B^\ell(b) \exp[s^\ell(a, b)], \quad [2]$$

defining the log-likelihood *link similarity score* $s^\ell(a, b)$. For binary links, this has a bilinear form

$$s^\ell(a, b) = \lambda_\ell ab + \sigma_A^\ell a + \sigma_B^\ell b + \text{const.}, \quad [3]$$

with the link match reward λ_ℓ depending on the evolutionary distance between the species. The additive constant is given by the normalization of the probability distributions in Eq. 2. For continuous links, we write the joint distribution as $q^\ell(a, b) = G(b|a) p_A^\ell(a)$, where $G(b|a)$ is the conditional distribution of link strengths b evolved from an initial strength a over the evolutionary distance between the two species compared. For short evolutionary distances and for a link evolution of the form 1, this distribution is well approximated by a Gaussian, $G(b|a) \sim \exp[-\lambda_\ell g[(a+b)/2](a-b)^2]$. For large evolutionary distances, it can be shown that the score $s^\ell(a, b) = G(b|a)/p_B^\ell(b)$ has again the asymptotic form 3. Given datasets of two networks A and B ,

the distribution $p_{A'}^{\ell}, p_{B'}^{\ell}$ can be estimated from the frequency of link strengths $a_{i'j'}$ respectively $b_{i'j'}$ in one species and q^{ℓ} from the frequency of link pairs $(a_{i'j'}, b_{i'j'})$ involving orthologous gene pairs (i, j) and (i', j') . Hence, the score function $s^{\ell}(a, b)$ defined by Eq. 2 can be inferred without specific assumptions on the underlying link dynamics. For the example discussed below, this empirical link score turns out to be in remarkable agreement with the form 3 predicted by the link diffusion model.

This scoring of individual link pairs is readily generalized to pairs of networks (A, B) with a given local alignment π . Assuming that aligned link pairs $(a_{i'j'}, b_{\pi(i)\pi(j)})$ follow the distribution $q^{\ell}(a, b)$ independently from each other and unaligned links $a_{i'j'}$ and $b_{i'j'}$ follow the distributions $p_{A'}^{\ell}(a)$ and $p_{B'}^{\ell}(b)$, respectively, we obtain the distribution of graph pairs for given π ,

$$Q^{\ell}(\mathbf{a}, \mathbf{b} | \pi) = P_{A'}^{\ell}(\mathbf{a}) P_{B'}^{\ell}(\mathbf{b}) \exp[S^{\ell}(\mathbf{a}, \mathbf{b}, \pi)], \quad [4]$$

where $P_{A'}^{\ell}(\mathbf{a}) = \prod_{i,j \in A} P_{A'}^{\ell}(a_{ij})$, $P_{B'}^{\ell}(\mathbf{b})$ has a similar product form, and the network link score $S^{\ell}(\mathbf{a}, \mathbf{b}, \pi)$ is a sum of local contributions $S_{i,\pi(i)\ell}$ of aligned node pairs,

$$S^{\ell}(\mathbf{a}, \mathbf{b}, \pi) = \sum_{i \in A} S_{i,\pi(i)\ell} = \sum_{i,j \in A} s^{\ell}(a_{ij}, b_{\pi(i)\pi(j)}), \quad [5]$$

as shown in Fig. 1b. For coexpression networks, there are correlations between links within one network. These occur because the number of independent measurements, d , is smaller than the number of genes N , and is taken into account by the overall scale of the link score (i.e., $\lambda_{\ell} \sim \sigma^{\ell} \sim d/N$).

The relative evolutionary conservation of a given pair (a, b) of aligned links within the network is measured by its excess link score

$$\Delta s^{\ell}(a, b) \equiv s^{\ell}(a, b) - (\langle s^{\ell}(a, b') \rangle_{b'} + \langle s^{\ell}(a', b) \rangle_{a'}) / 2, \quad [6]$$

i.e., the difference of its link score and the average over all aligned link pairs with either strength a fixed, $\langle s^{\ell}(a, b') \rangle_{b'} \equiv \int db' G(b'|a) s^{\ell}(a, b')$, or strength b fixed. The relative conservation of link patterns between a pair of aligned nodes i , $\pi(i)$ is then given by the local excess link score

$$\Delta S_{i,\pi(i)\ell}^{\ell} = \sum_{i',j' \in A} \Delta s^{\ell}(a_{i'j'}, b_{\pi(i)\pi(j')}). \quad [7]$$

These measures will become important for the identification of network clusters and their evolutionary conservation.

Node Dynamics and Node Score. The pairwise similarity between genes in networks A and B is given by a matrix Θ , whose entries θ_{ij} quantify, for example, the overall sequence similarity between the gene sequences $i \in A$ and $j \in B$ or a biochemical similarity between the corresponding proteins. The sequence similarity between functionally related genes decays over time because of local mutations, but it is also affected by large-scale genomic events such as gene duplications, gene loss, or recruitment of new genes into a functional context. Because of these processes, both networks contain a fraction of nodes with little or no significant sequence similarity to any node in the other network, which should nevertheless be included in the alignment if their local link score suggests significant functional cross-species relationships. Moreover, *functional swaps* between genes induce functional correlations between genes that are unrelated by sequence and, at the same time, reduce correlations between other genes despite their sequence similarity. A prominent example is nonorthologous gene displacements (13–15). It is these processes that cause the network alignment to deviate for some nodes from a map based only on sequence homology. Functional swaps can be regarded as part of the link evolution, which in coexpression networks leads to *coherent* link changes at the affected nodes. However, these swaps are not captured by the independent link dynamics discussed above. Hence, we include

them here as a separate type of process with its own evolutionary rate.

The resulting statistics of node similarity can be described by the distribution of pairwise similarity coefficients between unaligned nodes, $p_0^n(\theta)$, between pairs of aligned nodes, $q_1^n(\theta)$, and between one aligned node and nodes other than its alignment partner, $q_2^n(\theta)$. Note that $p_0^n(\theta)$ does not simply describe uncorrelated sequences: significant sequence similarity may exist between genes that are not aligned because of their link mismatch, because functional changes can lead to a rapid divergence of links, for example, in the formation of a pseudogene.

The log-likelihood *node similarity scores* $s_1^n(\theta)$ and $s_2^n(\theta)$, which are defined by

$$q_1^n(\theta) = p_0^n(\theta) \exp[s_1^n(\theta)], \quad q_2^n(\theta) = p_0^n(\theta) \exp[s_2^n(\theta)], \quad [8]$$

quantify the dependence of the alignment on node similarity. Assuming that the coefficients θ_{ij} are drawn independently from these distributions, we obtain the distribution of node similarity for a pair of networks A and B with a given alignment π ,

$$Q^n(\Theta | \pi) = P_0^n(\Theta) \exp[S^n(\Theta, \pi)], \quad [9]$$

where $P_0^n(\Theta) = \prod_{i,j} p_0^n(\theta_{ij})$ and the network node score $S^n(\Theta, \pi)$ is again a sum of local contributions $s_1^n(\theta_{ij})$ and $s_2^n(\theta_{ij})$. In this article, we use a simple binary approximation of node similarity: two genes are counted as orthologous ($\theta_{ij} = 1$) if they appear as putative orthologs in the Ensembl database (20), and otherwise not ($\theta_{ij} = 0$). Each node may have several such putative orthologs. The three distributions in Eq. 8 are then fully determined by three model parameters, $p_0^n(\theta) \sim \exp[\zeta_0 \theta]$, $q_1^n(\theta) \sim \exp[(\zeta_0 + \lambda_n) \theta]$, and $q_2^n(\theta) \sim \exp[(\zeta_0 + \lambda'_n) \theta]$, which in turn depend on the rates of the node dynamics and on the evolutionary distance between the species. A short calculation shows that the node score 9 takes the form

$$S^n(\Theta, \pi) = \sum_{i \in A} (S_{i\pi(i)}^n + \mu). \quad [10]$$

Here the local *node similarity score*

$$S_{i\pi(i)}^n = \begin{cases} 0 & \text{if } \sum_{j \in B} \theta_{ij} = \sum_{i' \in A} \theta_{i'\pi(i)} = 0, \\ \lambda_n & \text{if } \theta_{i\pi(i)} = 1, \\ \lambda'_n & \text{otherwise} \end{cases} \quad [11]$$

measures the overlap of alignment π and homology map Θ as shown in Fig. 1c, and the “chemical potential” $\mu(\lambda_n, \lambda'_n, \zeta_0)$ implicitly determines the overall number of nodes in the alignment (for details, see the *Supporting Text*, which is published as supporting information on the PNAS web site). For large μ , the highest scores occur in *global alignments* between the networks A and B , which involve all nodes of the smaller network. This behavior is appropriate if the evolution of links and nodes maintains for all nodes some functional relationship within the network. In the case of this study, link and node dynamics destroy significant correlations for some nodes. We obtain local alignments with chemical potential $\mu < 0$, which exclude some nodes of both networks.

Hidden Markov Model and Bayesian Analysis. We can now combine the distributions Q^{ℓ} and Q^n into a probabilistic model for link and node similarity, which produces the observable data, i.e., pairs of networks with adjacency matrices \mathbf{a}, \mathbf{b} and node similarity matrix Θ , for a given alignment π and for given model parameters $m = (s^{\ell}, \lambda_n, \lambda'_n, \zeta_0)$ in Eqs. 5 and 10. The combined model is given by the probability distribution

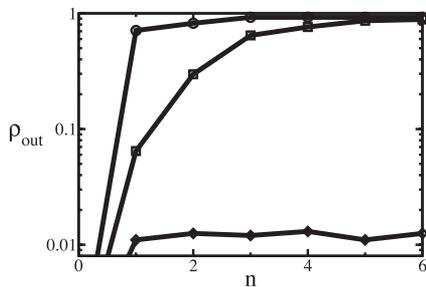


Fig. 2. Performance characteristic of the alignment algorithm. The fraction ρ_{out} of correctly aligned nodes is plotted against the number of iterative steps n for fractions $\rho_{in} = 0.01$ (\diamond), 0.02 (\square), or 0.5 (\circ) of the node similarity given as input. Typically the algorithm converges after about five iterations. There is a switch from low to high alignment quality ($\rho_{out} > 0.9$) at a threshold value $\rho_c \approx 0.02$.

$$Q(\mathbf{a}, \mathbf{b}, \Theta | \pi, m) = Q^\ell(\mathbf{a}, \mathbf{b}, \Theta | \pi, m) Q^n(\Theta | \pi, m) \\ = \exp[S(\mathbf{a}, \mathbf{b}, \Theta, \pi, m)] P_A^\ell(\mathbf{a}) P_B^\ell(\mathbf{b}) P_0^n(\Theta, \zeta_0) \quad [12]$$

with the alignment score function

$$S(\mathbf{a}, \mathbf{b}, \Theta, \pi, m) = S^\ell(\mathbf{a}, \mathbf{b}, \pi, m) + S^n(\Theta, \pi, m). \quad [13]$$

Eqs. 12 and 13 are at the heart of our scoring procedure: they provide a probabilistic rationale for the cross-species analysis of network data by link and node similarity. The model parameters m , which determine the relative weight of link and node score, and the alignment π are “hidden” variables, which can be inferred by a standard Bayesian analysis. We write their posterior probability, i.e., the conditional probability of the hidden variables for given data $\mathbf{a}, \mathbf{b}, \Theta$, in the form

$$Q(\pi, m | \mathbf{a}, \mathbf{b}, \Theta) = \frac{Q(\mathbf{a}, \mathbf{b}, \Theta | \pi, m) P(\pi, m)}{\sum_{\pi, m} Q(\mathbf{a}, \mathbf{b}, \Theta | \pi, m) P(\pi, m)} \quad [14]$$

and assume the prior probability $P(\pi, m)$ to be flat. Dropping the terms independent of π and m , we obtain the optimal local alignment π^* by maximizing the posterior probability $Q(\pi | \mathbf{a}, \mathbf{b}, \Theta) \sim \sum_m Q(\mathbf{a}, \mathbf{b}, \Theta | \pi, m)$ and similarly the optimal scoring parameters m^* by maximizing $Q(m | \mathbf{a}, \mathbf{b}, \Theta) \sim \sum_\pi Q(\mathbf{a}, \mathbf{b}, \Theta | \pi, m)$. In a Viterbi approximation, π^* and m^* can be inferred jointly by maximizing $Q(\mathbf{a}, \mathbf{b}, \Theta | \pi, m)$. This amounts to determining the optimal null model parameter ζ_0 and maximizing the combined score $S(\mathbf{a}, \mathbf{b}, \Theta, \pi, m)$. Details are given in *Supporting Text*.

Alignment Algorithm. Our algorithm for maximizing the score is based on a mapping to a generalized quadratic assignment problem, which is solved by an iterative heuristic similar to that in ref. 19 with

running times of order N^3 (21) (for details, see *Supporting Text*). To quantify the performance of the algorithm for coexpression networks, we have used a human microarray dataset (22), consisting of expression measurements of various tissues. We randomly partitioned the experiments into two equally large subsets, and thus obtained two “mirror copies” of the expression correlation network in one species. The nodes in the two networks are identical and their links differ only by experimental noise. The correct alignment of these two copies is trivial, $\pi(i) = i$. A fraction ρ_{in} of correctly aligned nodes with randomly chosen indices i is given as input for the algorithm by specifying the corresponding node similarity coefficients $\theta_{ij} = \delta_{ij}$, the remaining node information is ignored ($\theta_{ij} = 0$). We then record the fraction of correctly aligned nodes $\rho_{out}(n)$ of the algorithm as a function of the number of iterations n (see Fig. 2). This performance characteristic shows a switch from low to high alignment quality at a threshold value $\rho_c \approx 0.02$. In the low-quality regime ($\rho_{in} < \rho_c$), the alignment contains for all n only the nodes given as input. In the high-quality regime ($\rho_{in} > \rho_c$), the iterations continuously improve the fraction of correctly aligned nodes, saturating at an accuracy $\rho_{out} > 0.9$ for large n . Of course, the threshold will be higher and the saturation accuracy lower for cross-species comparisons, where the networks differ by evolutionary changes and by larger experimental variation. Similarly, the threshold rises if the network is randomly diluted (to $\rho_c \approx 0.2$ when 95% of all links have been set to zero).

Results

Aligning Human/Mouse Expression Data. The coexpression networks were constructed from the expression data of Su *et al.* (22) obtained from 79 tissues in humans, 61 tissues in mice, and a set of biological and technical replicates of the same size. Similar experimental protocols were used in both species, making the data particularly suitable for cross-species comparison. Our networks A (human) and B (mouse) of size $N_A = N_B = 2,065$ contain genes that are expressed in all samples and show a low variance of expression levels across samples in both species (housekeeping genes), as well as genes having a high expression similarity with at least one such housekeeping gene. The link strength $a_{ii'}$ is defined as the Spearman correlation between the expression levels of the human genes i and i' across all tissues, and similarly $b_{jj'}$ in mouse. Both networks have a broad distribution of link values; the distribution $p_A^\ell(a)$ in human is shown in Fig. 3a. To determine the link scoring function $s^\ell(a, b)$, we look at all human gene pairs (i, i') that have homologs (j, j') in mouse and compute the distribution of link pairs $a = a_{ii'}$ and $b = b_{jj'}$. The optimal alignment π (along with the optimal node model parameters $\lambda_n, \lambda'_n, \zeta_0$) is then inferred by likelihood maximization as described above; it consists of 1,956 genes.

The overall cross-species variation of expression is given by the root mean square difference $\Delta_\ell \equiv \sqrt{\langle (a - b)^2 \rangle}$ (with the brackets $\langle \dots \rangle$ indicating the average over all aligned link pairs a, b), we find $\Delta_\ell = 0.33$. To separate this variation into evolutionary changes and

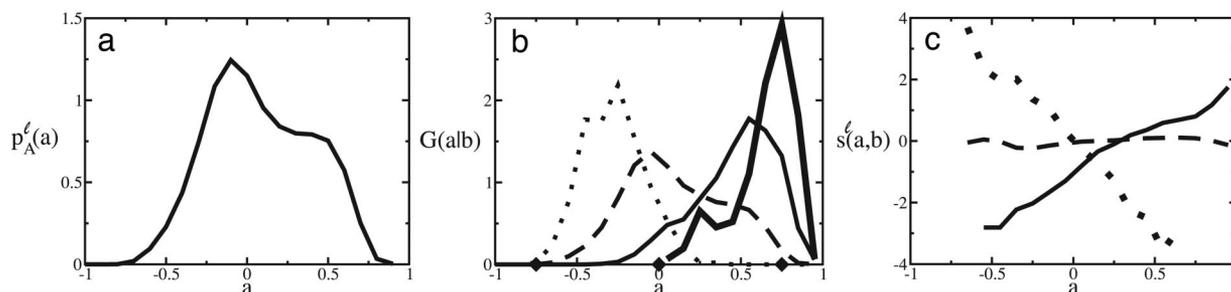


Fig. 3. Evolution of coexpression links between human and mouse. (a) The distribution of $p_A^\ell(a)$ of link strengths in human. (b) The conditional distribution of link strengths in human, $G(a|b)$, plotted against a for the values $b = -0.75$ (dotted line), 0 (dashed line), and 0.75 (solid line) in mouse. The heavy solid line shows the conditional distribution $G(a|b = 0.75)$ restricted to links within expression clusters; see text. (c) The empirical link scoring function $s^\ell(a, b)$ for $b = -0.75$ (dotted line), 0 (dashed line), and 0.75 (solid line).

sampling noise, we again construct coexpression networks from a randomly chosen subset containing half the expression measurements from either organism and obtain $\Delta_\ell = 0.35$, i.e., sampling contributes only a small fraction to Δ_ℓ .

To trace the link evolution between the networks in more detail, we look at the conditional distribution $G(a|b)$ of correlation values $a_{i'} = a$ in humans given a certain correlation value $b_{\pi(i)\pi(i')} = b$ in mouse, which is shown in Fig. 3b as a function of a for three different values of b . As expected, the variance of $G(a|b)$ is largest for weak correlations and less for strong positive or negative correlations. The resulting link scoring function $s^\ell(a, b) = \log[G(a|b)/p_A^\ell(a)]$ is a linear function of a with the slope determined by b ; see Eq. 2 and Fig. 3c.

Conserved Network Patterns. Coexpression networks are not homogeneous (1). Instead, they are organized in clusters of genes that have similar expression profiles. In the mouse network, we call a gene j *clustered* if it has a correlation $b_{ji} > 0.8$ with more than 15 other genes (the average number of links $b > 0.8$ is ≈ 1 per gene). With this definition, there are 40 clustered genes in the network B (little of the following depends on the exact thresholds chosen). The thick line in Fig. 3b shows the conditional distribution $G(a|b = 0.75)$ restricted to links $b = b_{ji}$ involving a clustered mouse gene j . The root mean change of the expression correlations is $\Delta_\ell = 0.22$. This is a reduction by a factor of 2, compared with the distribution $G(a|b = 0.75)$ for all genes. This reduced change of expression correlation for clustered genes translates into a local excess link score (Eq. 7) of $\Delta S^\ell \sim 10$ per gene. This finding suggests that clustered genes have more strongly conserved expression patterns than genes that are not part of clusters (see also ref. 1). Fig. 4(a) shows the link evolution between a set of clustered genes (arranged in a circle) and a randomly chosen set of genes outside this cluster (arranged in a straight line). The link intensity encodes the correlation strength a in human, the color its evolutionary conservation as measured by the excess link score $\Delta S^\ell(a, b)$. Intracuster links are, on average, stronger (i.e., more intense) and at the same time more conserved (i.e., contain more blue) than links with genes external to the cluster. The genes contained in this cluster are involved in the control of transcription and code for constituents of the ribosome; their full list is given as Tables 1–3, which are published as supporting information on the PNAS web site.

Correlations Between Link and Node Similarity. Fig. 5 shows an overall correlation between cross-species sequence similarity quantified by the score of the best nucleotide BLAST hit (23) and link similarity measured by the excess link score ΔS^ℓ . Gene pairs with a high sequence score also have a bias toward high link similarity. However, the converse is not true: most of the gene pairs with strongly conserved expression patterns have only average sequence similarity. An example is the gene cluster discussed above (marked by gray diamonds in Fig. 5), which has a significant excess link score $\Delta S^\ell \sim 10$ and a sequence score of 440 per gene, which is not significantly above the network average of 394.

Network-Based Gene Annotations. Network alignment as a putative functional map differs from the homology map of individual genes: there are genes without an (easily detected) homologous partner in the other network. These genes are aligned solely on the basis of their link score. Although our dataset is centered on housekeeping genes and may be biased against such cases, the maximum-likelihood alignment contains significant cases of such link-based alignments, which are reported in Tables 1–3 (and marked by gray circles in Fig. 5).

(i) Human *ORIC1* is aligned to mouse *Olfir836* with a local link score $S^\ell = 16.1$ exceeding the average value 6.7 between orthologs. A functional relationship between these genes is quite plausible: Not only are both genes involved in olfactory receptor activity (24), they also have two protein domains in common and belong to the

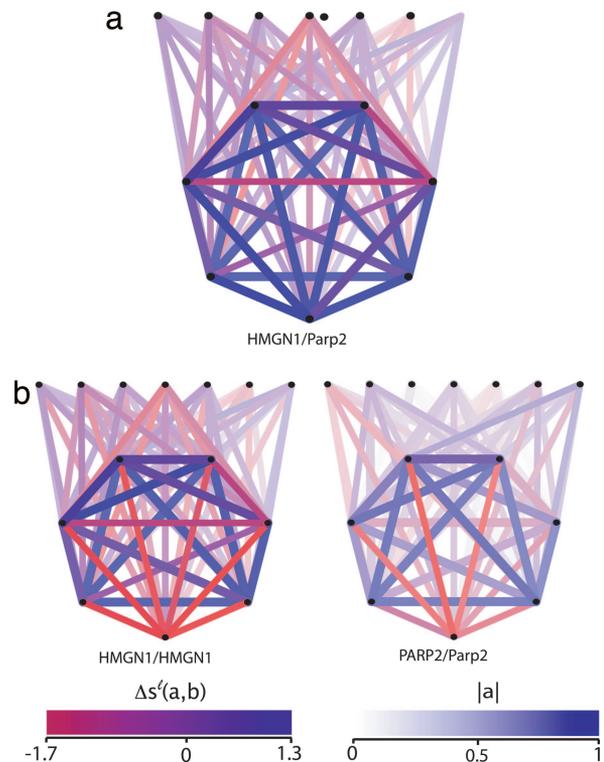


Fig. 4. Evolutionary conservation of gene clusters. (a) Seven genes from a cluster of coexpressed genes together with seven random genes outside this cluster (straight line). Each node represents a pair of aligned genes in human and mouse. The intensity and color shading of a link encode the correlation value a in human and its relative evolutionary conservation between the two species (see color bars). This cluster contains the nonorthologous aligned gene pair human *HMGN1* (high-mobility group nucleosome-binding domain 1)/mouse *Parp2* [poly(ADP-ribose) polymerase], predicting a nonorthologous gene displacement. (b) The same cluster, but with human *HMGN1* “falsely” aligned to its ortholog mouse *HMGN1* (Left), and human *PARP2* aligned to its ortholog mouse *Parp2* (Right, with the intensity encoding the correlation in mouse). This mismatch shows the poor expression similarity for this pair of genes.

same gene family. However, their overall DNA sequence identity is below 60%, compared with a typical value of 85% between orthologs in human and mouse. Most likely these genes are distant

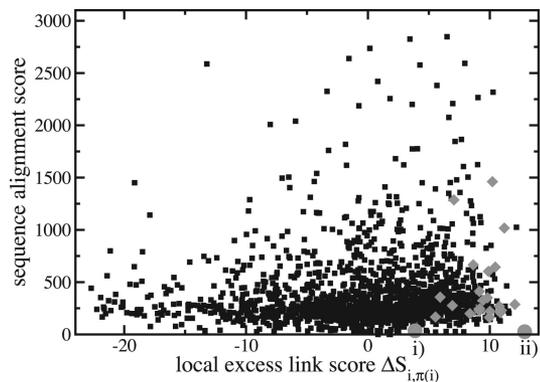


Fig. 5. Node versus link evolution. For aligned pairs of genes ($i, \pi(i)$), the nucleotide BLAST score with standard parameters (vertical axis) is plotted against the excess link score $\Delta S_{i,\pi(i)}^\ell$ (horizontal axis). Genes in the cluster shown as gray diamonds are distinguished by high link similarity, but most of them show no enhanced BLAST score. The gene pairs i and ii , aligned solely on the basis of the link score (see text), are indicated by gray circles.

orthologs, predating the human–mouse split. This is an example where functional constraints maintain a high level of conservation at the network level, but not at the sequence level.

(ii) In the case human *HMGNI*/mouse *Parp2*, both genes have orthologs but the network alignment does not match the orthology map. As shown in Fig. 4*a*, the human gene *HMGNI* is part of a gene cluster, and the alignment to mouse *Parp2* (with $S^{\ell} = 25.1$) respects the evolutionary conservation of that cluster. The “false” alignment human *HMGNI*/mouse *HMGNI* respects orthology but produces a link mismatch ($S^{\ell} = -12.4$) because of the poor expression similarity of mouse *HMGNI* with the other genes of the cluster; see Fig. 4*b*). Human *HMGNI* is known to be involved in chromatin modulation and to act as an RNA polymerase II transcription factor, in particular through altering the accessibility of regulatory DNA. The network alignment predicts a similar role of *Parp2* in mouse, which is distinct from its known function in the poly(ADP-ribosylation) of nuclear proteins. This prediction is consistent with a recent experimental study (25) inhibiting the members of the *Parp* gene family in mouse. The authors conclude that “in addition to known functions of poly(ADP-ribosylation), some so far unrecognized, nonredundant functions may also exist”; specifically the chromatin remodeling involved in gene expression changes during development.

Discussion

Alignment Provides a Quantitative Measure of Network Divergence.

We have developed a probabilistic alignment procedure for biological networks based on their link and node similarity. Both components of similarity are important, i.e., a network alignment differs, in general, both from a mere matching of link patterns and from a mere node homology map. To the extent that significant sequence homology is present, it clearly introduces a bias for the functional association of genes across organisms, and hence for the alignment. It is this bias that makes the problem computationally tractable: although there is probably no formal solution by a polynomial-time algorithm, biological network alignment allows for more efficient heuristics than generic pattern matching. (We have discussed here an alignment of $\approx 2,000$ genes, but ongoing studies suggest the method can be scaled up to genome-wide cross-species comparisons of vertebrates.) On the other hand, the homology relations are not completely respected even between relatively close species: network alignment thus predicts a deviation of functional evolution from sequence evolution for some genes. Assessing the statistical significance of such functional swaps requires tuning the relative weight of link and node similarity in a consistent way, which is done here by a Bayesian inference from the datasets.

Cluster Conservation and Selection. There are important differences in the population genetics of sequences and networks. Sequence

divergence has an approximate molecular clock of synonymous nucleotide changes, which can be described approximately by neutral evolution. Adaptive changes can be quantified relative to neutral evolution. For interaction networks, the relative weights of neutral evolution, negative selection, and positive selection are far less clear. Indeed, the role of selection in the evolution of expression patterns is currently under debate (26, 27). Even the direction of evolution may not be as predominantly divergent as for sequences: the selection for a given function may lead to convergent evolution of network structures. Nevertheless, there is some regularity in the evolution of expression patterns: genes that are part of a strongly correlated cluster in one species have a significantly reduced cross-species variation of their expression profile; this conservation is quantified by a typical excess link score ΔS^{ℓ} of order 10 per gene. Selection for functionality is indeed a possible explanation. However, as the example of Fig. 4 shows, selection in a network can be rather complex: conservation of a gene cluster as a whole could be attributed to purifying selection at the level of network interactions, but this does not exclude positive selection leading to functional swaps at the level of network constituents.

Network-Based Prediction of Gene Function. Given a cross-species alignment of gene networks, we can quantify link and node evolution. For our cross-species analysis between humans and mice, the correlations between these two modes are shown in Fig. 5. Although high sequence similarity predicts high link conservation, most of the gene pairs with high link conservation have only average sequence similarity. Hence, the network alignment contains functional information beyond the corresponding sequence alignment: it detects evolutionary conservation that is not discernible by a comparison of overall similarity between sequences. Identifying genes with conserved expression patterns will also aid the cross-species analysis of regulatory binding sites, where a rapid turnover of binding sites despite the conservation of expression patterns has been found (28). Extreme cases of mismatch between link and node evolution are gene pairs with significantly similar interaction patterns but with no significant sequence similarity at all. This mismatch can be due to long-term sequence evolution between orthologous genes, which randomizes their sequence similarity, whereas their functional roles are more conserved. It may also arise from link dynamics leading to link similarities between genes that are completely uncorrelated at the sequence level (13–15). In our alignment of coexpression networks, we find evidence for both processes. Thus, the alignment leads to functional predictions on the basis of network similarity alone, in cases where a functional annotation is known for one of the aligned genes.

We thank Terence Hwa, Daniel Barker, and Diethard Tautz for discussions. This work was supported through Deutsche Forschungsgemeinschaft Grants SFB/TR 12, SFB 680, and BE 2478/2-1.

1. Stuart, J., Segal, E., Koller, D. & Kim, S. (2003) *Science* **302**, 249–255.
2. Bergmann, S., Ihmels, J. & Barkai, N. (2004) *PLoS Biol.* **2**, e9.
3. Sharan, R., Suthram, S., Kelley, R., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. & Ideker, T. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979.
4. Sharan, R., Ideker, R., Kelley, B., Shamir, R. & Karp, R. (2005) *J. Comput. Biol.* **12**, 835–846.
5. Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002) *Nat. Genet.* **31**, 64–68.
6. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002) *Science* **298**, 824–827.
7. Berg, J. & Lässig, M. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 14689–14694.
8. Trusina, A., Sneppen, K., Dodd, I., Shearwin, K. & Egan, J. (2005) *PLoS Comput. Biol.* **7**, e74.
9. Heymans, M. & Singh, A. (2003) *Bioinformatics* **19**, 138–146.
10. Kelley, B., Sharan, R., Karp, R., Sittler, T., Root, D., Stockwell, B. & Ideker, T. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 11394–11399.
11. Kelley, B., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. & Ideker, T. (2004) *Nucleic Acids Res.* **32**, W83–W88.
12. Pinter, R., Rokhlenko, O., Yeger-Lotem, E. & Zvi-Ukelson, M. (2005) *Bioinformatics* **21**, 3401–3408.
13. Koonin, E., Mushegian, A. & Bork, P. (1996) *Trends Genet.* **12**, 334–336.
14. Galperin, M. Y. & Koonin, E. V. (1998) *In Silico Biol.* **1**, 55–56.
15. Morett, E., Korbel, J., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B. & Bork, P. (2003) *Nat. Biotechnol.* **21**, 790–795.
16. Needleman, S. & Wunsch, C. (1970) *J. Mol. Biol.* **48**, 443–453.
17. Smith, M. S. & Waterman, T. F. (1981) *J. Mol. Biol.* **147**, 195–197.
18. Papadimitriou, C. (1995) *Computational Complexity* (Addison-Wesley, Reading, MA).
19. Tsafir, D., Tsafir, I., Ein-Dor, L., Zuk, O., Notterman, D. A. & Domany, E. (2005) *Bioinformatics* **21**, 2301–2308.
20. Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., et al. (2005) *Nucleic Acids Res.* **33**, D447–D453.
21. Jonker, R. & Volgenant, A. (1987) *Computing* **38**, 325–340.
22. Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067.
23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
24. The Gene Ontology Consortium (2000) *Nat. Genet.* **25**, 25–29.
25. Imamura, T., Anh, T. M., Thenevin, C. & Paldi, A. (2004) *BMC Mol. Biol.* **5**, 4.
26. Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. & Pääbo, S. (2004) *PLoS Biol.* **2**, E132.
27. Yanai, I., Korbel, J. O., Boue, S., McWeeney, S., Bork, P. & Lercher, M. J. (2006) *Trends Genet.* **22**, 132–138.
28. Tanay, A., Regev, A. & Shamir, R. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 7203–7208.