An Approach for Determining and Measuring Network Hierarchy: Application to Comparing the Phosphorylome and the Regulome

Chao Cheng[1,2,3*], Erik Andrews[1], Koon-Kiu Yan[4], Matthew Ung[1], Daifeng Wang[4], Mark Gerstein [4,5,6*]

[1]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA.
[2]Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA.
[3]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA.
[4]Program in Computational Biology and Bioinformatics, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
[5]Department of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA
[6]Department of Computer Science, Yale University, 260 Whitney Avenue, New Haven, CT 06520, USA

*Corresponding authors

Abstract

Many biological networks naturally form a hierarchy with a preponderance of downward information flow. In this study, we define a score to quantify the degree of hierarchy in a network and develop a simulated-annealing algorithm to maximize the hierarchical score globally over a network. We apply it to compare the regulatory and phosphorylation networks, and find that the phosphorylome is more hierarchical than the regulome. Furthermore, we determined the hierarchical structure of yeast phosphorylome and investigated the correlation between hierarchy and kinase properties.

Introduction

Networks have been used as universal frameworks to represent many complex systems including the World Wide Web [1], social interactions [2], literature citation relationships [3], and biological processes [4-6]. Based on the attribute of edges, networks can be subdivided into two categories: undirected and directed networks. In an undirected network there is no distinction between the two vertices associated with each edge, whereas in a directed network all edges are directed from one vertex to another. The asymmetric nature of edges in a directed network causes topological differences of nodes, resulting in a hierarchical structure: some function as top regulators, while others function as downstream effectors.

Owning to the development of large-scale experimental techniques, many biological networks have been produced. These include protein-protein interaction networks and genetic interaction networks, etc [7-12]. Among them, the gene regulatory network (referred to as the regulome) and the protein phosphorylation network (referred to as the phosphorylome) are the two best-studied directed networks [10, 11]. The regulome captures the transcriptional regulatory interactions of transcription factors (TFs) with their target genes. The techniques to systematically identify TF-DNA interactions include the bacterial one-hybrid system [13], the yeast one-hybrid system [14], and chromatin immunoprecipitation followed by microarray (ChIP-chip) [15] or parallel sequencing (ChIP-seq) [16]. In particular, ChIP-chip and ChIP-seq have been used to determine the target genes of a large number of TFs in recent years, and will produce more data in the near future. In particular, in yeast Harbison et al. have performed ChIP-chip experiments to identify target genes of 203 proteins, which represent

nearly all of the DNA-binding transcriptional regulators encoded in the yeast genome [10]. In human, the Encyclopedia of DNA Elements (ENCODE) project has determined the genomic binding sites of more than 120 TFs [17]. Meanwhile, the interactions between kinases, phosphotases, and their substrates can be identified by protein chip [11] or mass spectrometry [18]. The latter technology is capable of providing precise phosphorylation sites. In particular, Ptacek et al has determined the in vitro substrates recognized by most yeast protein kinases [11]. The availability of these datasets enables us to construct regulomes and phosphorylomes and investigate the regulatory mechanisms of TFs and kinases on a systems level.

Since regulome and phosphorylome are directed networks, it is of particular interest to examine whether they harbor a hierarchical structure (TF/kinase nodes function at different levels) and, if so, how that hierarchy is organized. Particularly, we have previously investigated the rewiring of the regulomes in E. coli and S. cerevisiae, and found that hierarchy, rather than connectivity, better reflects the importance of regulators [19]. For the regulomes, the hierarchy properties have been explored in several studies [17, 20-23]. In these studies, the authors inferred the hierarchical structure of regulomes and examined the correlation between the hierarchy and TF features. For example, Jothi et al demonstrated that top-level TFs in the yeast regulome are more likely to be essential and are more conserved across species [21]. These studies provide critical insights into the regulatory mechanisms of TFs during transcription regulation. On the other hand, the phosphorylation network has not been investigated from a hierarchical perspective.

Several algorithms have been proposed to infer the hierarchical structure of directed networks [20, 21, 24-27], including the leaf removal (LF) algorithm, the breadth-first-search (BFS) method and the vertex sort (VS) algorithm. These algorithms have been applied to the regulome and revealed new insights on hierarchical organization of TFs during transcriptional regulation. Despite their effectiveness, they have several limitations and do not address some important issues about hierarchical networks. The LF algorithm determines the hierarchical structure by removing leafs iteratively, and as a consequence it cannot be applied to a directed network with cycles. Similarly, for networks with cycles the BFS method has to break cycles before assigning hierarchical levels to nodes [20]. In addition, these two methods do not allow ambiguity of hierarchy which is usually the case in many networks. The VS algorithm proposed by Jothi et al is capable of overcoming these shortcomings. However, it can only assign ambiguous nodes to an interval of possible levels without providing the probability of them in each level [21].

Moreover, these methods have not addressed several important questions related to hierarchical networks: How to quantify the degree of hierarchical structure for a given network? How to estimate significance of hierarchical structure of a directed network? How to compare the degree of hierarchy of two directed networks? Importantly, can every node be assigned to a specific level with the same confidence? If not, how can we know which nodes are more confident than the others? For those ambiguous nodes, what are the probabilities of them to be assigned to each level?

To address these questions, we define a metric to quantify the degree of hierarchy for a given hierarchical network, and then propose a new method called hierarchical score maximization (HSM) to infer the hierarchy of a directed network. First, we apply the algorithm to a military command network. The results demonstrate its effectiveness in precisely determining the network's hierarchy. Second, we apply the algorithm to eight directed networks including biological networks, social networks, and ecological networks. We compare these networks in terms of their degrees of hierarchy and the results suggest that phosphorylomes are more hierarchical than transcriptional regulatory networks. We compare the hierarchical structure of the yeast regulome determined using the HSM algorithm with those from previous algorithms. Finally, we investigate the hierarchical structure of the yeast phosphorylome and relate kinases in different levels with different genomic features.

Results

Construction of hierarchy by simulated annealing

To infer the hierarchical structure of a directed network, we start by defining a score to quantify the degree of hierarchy. For a network with a specified hierarchical topology (i.e., every node is assigned to a specific hierarchy level), there are in general three types of edges: downward interactions (pointing from higher-level to lower-level nodes), upward interactions (pointing from lower-level to higher -level nodes) and horizontal interactions (between nodes in the same level). We thus define the hierarchy score (HS) as the ratio of the number of downward interactions (N_d) to the number upward interactions (N_u) balanced by the number of horizontal interactions (N_h) (see "Methods" for details) (Fig. 1A). Based on this definition, we infer the hierarchical structure of a directed network as the one that achieves the maximum hierarchy score. Specifically, a simulated annealing algorithm is used to continuously adjust the hierarchical structure until the hierarchy score is maximized (Fig. 1B). The HS for two hierarchical networks with different numbers of levels are in general not directly comparable. To address this issue, we therefore revise the HS into a new metric called the corrected hierarchy score (CHS), which quantifies the enrichment in downward flow relative to expectation (see "Methods" for details). Finally, we define a p-value for how likely one could get such a hierarchical structure randomly.

In principle, the optimum hierarchical structure for a directed network may not be unique due to the existence of loops. Some nodes can be assigned to different levels without significant change of hierarchy score. For this reason, it is more reasonable and more informative to represent the hierarchical structure as a probabilistic model, in which a node may be assigned to multiple levels with different probabilities. To estimate these probabilities, for a directed network we performed the simulated annealing procedure for 1000 times (k=1000), which results in a probabilistic hierarchical network (Fig. 1C). Accordingly, we define a score called the probabilistic hierarchy score (PHS) to more accurately quantify the hierarchy underlying a directed network (see "Methods" for details). Typically, most of the nodes have a favored level to which the node is assigned with a significantly higher probability than the other levels. We thus can obtain a determined hierarchical structure by assigning each node to its most probable level (Fig. 1C).

Application of the HSM algorithm to a military command network

To show the effectiveness of our method we apply it to a military command network, which is a directed acyclic graph (DAG) with a perfect hierarchical structure. Since there are no loops in the network, the hierarchy levels of each node can be deterministically assigned (Fig. 2A). We then apply the HSM algorithm to the network, specifying different number of levels L=2, 3,…, 8. As shown in Fig. 2B, the hierarchical structure is precisely inferred when the correct number of levels (L=5) is specified. All of the nodes are assigned to the right levels with 100% certainty. Meanwhile the largest HS, CHS and PHS were obtained when L is set to 5. In practice, we do not have the prior knowledge about the number of hierarchical levels. Our simulation results suggest that to determine the number of levels we can try different L values and select the most reasonable one based on the CHS and PHS of the resulting hierarchical networks.

To show that the HS can quantify the degree of hierarchy of a directed network we perturb the original network in Fig. 1A by introducing a number of upward edges. We randomly introduce a number of (n) upward edges to perturb the hierarchical structure of the network. At each n, we repeat this procedure for many times and re-determine the hierarchical structure

of the perturbed networks. As shown in Fig. 2B, with the increase numbers of perturbations the hierarchy scores of the perturbed networks decrease asymptotically, indicating that the HS is an effective measurement for quantifying the degree of hierarchy of directed networks.

Hierarchical scores of several directed networks

We next apply the HSM algorithm to calculate the degree of hierarchy of eight different directed networks, including five biological networks (yeast regulatory network, human regulatory network, yeast phosphorylation network, human phosphorylation network and worm neural network), one ecological network (food web), one social network (political blogs) and one computer network (P2P file sharing network) (see "Methods" for information about these networks). We can evaluate the performance of the HSM algorithm, since we have an intuitive sense of the degree of hierarchy of these non-molecular networks.

In Table 1, we summarize the topological properties of these eight networks being sorted in the increasing order of CHSs. The political blog network contains hyperlinks between weblogs on US politics being recorded in 2005 [3]. The weblogs refer to each other by hyperlinks largely in a non-hierarchical manner, and consistently, we observe a relatively low hierarchical structure of it (CHS=3.2). In contrast, the food web network typically is known to have a pyramidal structure: the number of predators at each level decreases significantly, so that a single top predator is supported by a much larger number of preys. Indeed, the food web network is more hierarchical, with a CHS of 6.4. In addition, the worm neural network is the least hierarchical one among these networks, consistent to our knowledge that neurons are not hierarchically but mutually connected with one another [28].

Our results reveal several interesting findings. First, in both human and yeast, the phosphorylome is more hierarchical than the regulome (Table 1 and Fig. 3), although all these four network show significant hierarchical structures compared to a random network ($P<2e-16$, see method for significance estimation) [[Ref1.2.2]]. This is seen with the corrected hierarchy scores (CHSs) for yeast regulome and human regulome of 3.9 and 5.6, respectively, in contrast to the CHSs for yeast phosphorylome and human phosphorylome of 13.4 and 14.0, respectively. Surprisingly, the phosphorylomes are even more hierarchical than the food web network. Strikingly, all previous hierarchical network studies have been focused on regulomes and overlooked the phosphorylomes [17, 20, 21]. Our findings suggest that more investigation into the hierarchical nature of phosphorylome is warranted. Second, the degrees of hierarchy for both regulome and phosphorylome are highly consistent between yeast and human, the two evolutionarily distant species.

Comparison with other hierarchy construction algorithms

To compare the HSM algorithm with other methods, we apply it to the yeast regulome which contains 580 regulatory interactions among 145 transcription factors. With the same dataset, Yu et al. have applied a BFS method to construct a four-level hierarchical network [20]; Jothi et al. have applied a vertex sort (VS) approach to obtain a hierarchical network with seven levels, and further merged them into three levels [21]. We execute the HSM algorithm and obtain hierarchical networks with 3, 4, …, 8 levels. According to the CHSs, the hierarchical network with four levels is the most appropriate one.

We compare the CHSs of the hierarchical networked inferred by different methods (Fig. 4A). As HSM is designed to maximize the hierarchical score it gives rise to networks with significantly higher CHSs than those by BFS and VS methods (Fig. 4A). The hierarchical networks inferred by the other two methods have much lower CHSs than the optimum score. Moreover, the hierarchical network inferred by the HSM algorithm shows the highest fraction of downward interactions with >70% of interactions pointing from higher to lower level TFs. This is contrast to BFS and VS where <50% of interactions are downward. Although there are

no upward interactions in the hierarchical network derived from the VS algorithm (L=3), it has more horizontal interactions than the HSM algorithm (Fig. 4B). A similar fraction of horizontal edges are observed in the seven-level hierarchical network inferred by the VS algorithm. [[Ref1.2.1]]

We next examine the properties of TFs in relation to the hierarchy inferred by the HSM algorithm. As shown in Fig. 4A, the hierarchical network for yeast regulome with four levels (L=4) achieves the highest CHS, but the CHS for the network with three levels (L=3) is just slightly lower. In order to simplify the downstream analysis and to facilitate the comparison with previous studies we focus our analysis on the one with three TF levels, with 42, 41 and 62 TFs at the top, middle and bottom levels, respectively (Suppl. Table S1). First, we compare the percentage of essential TFs in the three levels. Our results indicate that higher level TFs are more likely to be essential if being deleted: 5 out of 42 top level TFs (12%) and 3 out of 41 middle level TFs (8%) are essential. In contrast, none of the 62 bottom level TFs is essential (P=0.01, Fisher exact test). In line with this, the TFs at the higher levels are more conserved during the evolution with the top level TFs tend to having a lower dN/dS ratio (calculated based on *S. cerevisae* versus *S. pombe* comparison) than the middle and bottom level TFs (P=0.008, Wilcoxon rank sum test). These results are consistent with those previously reported in Jothi et al. [21]. Second, we examine the degrees of the TFs at different levels in the physical interaction and genetic interaction networks. We find that TFs in higher levels (T+M) have significantly more physical interactions (P=0.0006, Wilcoxon rank sum test) than those in the bottom level, consisting with our observation in human regulome [17]. A similar trend is observed for genetic interactions, but it does not pass the significance threshold (P>0.05). Third, we compare the TFs at different levels on their dynamic properties, including their abundance and stability at both the mRNA and protein level, and their protein expression noise. The results indicate that the top-level TFs are more stable than middle- and bottom- level TFs (P=0.03, Wilcoxon rank sum test) (Fig. 4C). Overall, our results highlight the critical roles played by the top-layer TFs, as also reported by Jothi et al using the VS algorithm [21]. These master regulators are highly conserved during evolution with higher essentiality rate. [[Ref1.2.3]]

Features of kinases at different levels

Our results suggest that the organization of phosphorylomes is more hierarchical than the regulomes. We infer the hierarchical structure of the yeast phosphorylome by using the HSM algorithm. This network is mainly based on protein chip experiments and contains 200 phosphorylation interactions among 94 different kinases [11, 29]. Again, for easy comparison we specify the number of hierarchical levels L=3, which results in 38 top-level, 33 middle-level and 23 bottom-level kinases (Suppl. Table S2).

We examine the cellular localization according to the Saccharomyces genome database, which are manually annotated based on previous literatures. Of the 94 kinases 35 localize only to the cytoplasm, 8 only to the nucleus, and 12 to both (the remaining 39 kinases are in other locations or localization unknown). Interestingly, the kinases in the middle level are more likely to localize in both nucleus and cytoplasm compared to the top and bottom level kinases (P=0.02, Fisher exact test, Fig 5A). Gene ontology analysis suggests that the top-level is enriched in trans-membrane proteins and stress-response proteins implying that the top-level kinases tend to be located in the cell membrane and respond to extracellular signals (Suppl. Table S3). In contrast, the middle-level is enriched in cell cycle related kinases.

We also relate the hierarchical structure of the yeast phosphorylome with a number of kinase properties (Fig. 5B). Basically, our findings are summarized as the following. (1) The bud/bud-neck located proteins are highly enriched in kinases of the middle and bottom levels with respect to the top level (P=0.002, Fisher exact test). Strikingly, none of the 38 top-level kinases is a bud/bud-neck protein. This may suggest that during yeast budding the top-level

kinases function mainly in the mother cells rather than enter the bud/bud-neck to perform as direct effectors. (2) The middle-level kinases show higher essential rate (18%) than the top (8%) and the bottom (8%) level kinases. (3) Kinases in the middle level have significantly more physical (P=0.05, Wilcoxon rank sum test) and genetic (P=0.02, Wilcoxon rank sum test) interaction partners. (4) The bottom-level kinases are significantly noisier in their protein abundance than kinases in the higher levels (P=0.006, Wilcoxon rank sum test). ~~Together, these findings suggest that middle-level kinases might play critical roles in signal transduction -- they seem coordinate the phosphorylation signal flow from the cytoplasm into and out of nucleus.~~

Collaboration of kinases in different levels

We next explore how kinases in the top, middle and bottom levels collaborate with one another, in terms of both inter-level (TM, MB, TB) and intra-level (TT, MM, BB) relationships. First, we examine the physical and genetic interactions between kinases at different levels. Our results show that physical interactions are significantly enriched in TB (between top level and bottom level kinases) and MB (between middle level and bottom level kinases), but depleted in the intra-level relationships (TT, MM and BB). The genetic interactions are significantly enriched in MB, and depleted in TT and TB relationships (Fig. 6A). This suggests that inter-level interactions between kinases, particularly between middle and bottom level kinases, are dominant over those intra-level interactions.

Second, we investigate kinase cooperativity. We define two kinases as being cooperative if they share a significantly large number of physical partners, genetic partners or phosphorylation substrates (Fig. 6B). We find that physical cooperation between kinases is enriched in TB, while genetic cooperation is enriched in MB relationships. Interestingly, cooperation is highly depleted between bottom level kinases suggesting that, as downstream effectors, they tend to phosphorylate different subsets of proteins to take specific effects. Finally, we further divide genetic interactions into positive and negative ones, and examine their enrichment or depletion between kinases. Positive and negative genetic interactions involve a pair of genes with mutations or deletions in each alone causes a minimal phenotype, but when combined in the same cell results in a less severe (positive) or a more severe (negative) fitness defect than expected under a given condition [30]. As shown in Fig. 6C, both positive and negative genetic interactions are significantly enriched in MB relationships.

Substrate of kinases at different levels

The network contains 200 inter-kinase phosphorylation interactions (one kinase phosphorylating another) and 6 auto-phosphorylation interactions (CKA2, TPK2, RAD53, PRP1, CDC7 and CDC15). Indeed, the auto-phosphorylation is over-represented in the network (P=0.02, see "Methods" for details). There are two feedback loops (TPK2 and TPK3, ELM2 and GIN4) involving two kinases in the network in which the two kinases phosphorylate each other. The feed-forward loop (FFL) network motif is highly enriched in yeast phosphorylome. We investigate the FFL motifs in the context of hierarchy. In a FFL with three nodes, one kinase phosphorylates another kinase and both target a third protein as substrate, which can be either a kinase or non-kinase. We enumerate all of three-node FFL motifs in the yeast phosphorylation data (including non-kinase substrates) and map the two kinases in these motifs to the hierarchical network. Each of the two kinases in a FFL motif can be from one of the three hierarchical levels (T, M and B), which results in nine combinations (TT, TM, TB, MT, MM, MB, BT, BM and BB). We count the number of FFL motifs for all the nine types and our results show that >90% FFL motifs involve downward interactions between kinases in the hierarchical networks (Fig. 7A, red bars). The TM type

FFL motif, in which a top-level kinase phosphorylates a middle-level kinase and both kinases share a target substrate, is significantly enriched.

We also examine and compare the functions of the substrate targets of kinases at different levels. The 38 top-level kinases target a total of 1,095 substrates; the 33 middle-level kinases target 998 substrates; and the 23 bottom-level kinases target 612 substrates. The substrate targets of the three levels highly overlap as shown in Fig. 7B. After filtering out the shared substrate targets, we identify 294 top-level, 228 middle-level and 159 bottom-level specific substrate targets. Gene ontology analysis indicates that the top-level specific substrates are enriched in gene categories involving in "protein kinase activity", "phosphorylation", and "phosphate metabolic process", etc (Suppl. Table S4). In another words, the top-level kinases are involved in the regulation of other phosphorylation-related proteins. In contrast, the middle- and bottom-level specific substrate targets are enriched in structural proteins, e.g. gene categories involving in "microtubule cytoskeleton," "structural molecule activity" and "macromolecular complex subunit organization". ~~This is consistent with these proteins' more probable functions as downstream, effectors.~~

Discussion

Global optimization versus local optimization

To determine the hierarchical structure of a directed network all of the previous methods applied a local optimization strategy. The leaf removal algorithm employs a bottom-up iterative procedure. It assigns all the leaf nodes (nodes with zero out-degree) to the bottom level, and then iteratively removes all the leaf nodes and the edges associated with them from the network to determine the next higher level [25]. The BFS method also starts by assigning the leaf nodes to the bottom level, and then performs a BFS to define the level of a non-bottom node as its shortest distance from a bottom one [20]. The VS algorithm identifies strongly connected components and collapses them to convert the network into a directed acyclic graph, and then applies the leaf removal method to determine hierarchical levels [21]. All of these algorithms attempt to infer the hierarchical structure by iteratively optimizing the local hierarchy starting from the bottom level nodes.

To determine the hierarchical structure of a directed network, the leaf removal and the BFS methods applied a local optimization strategy. The leaf removal algorithm employs a bottom-up iterative procedure. It assigns all the leaf nodes (nodes with zero out-degree) to the bottom level, and then iteratively removes all the leaf nodes and the edges associated with them from the network to determine the next higher level [25]. The BFS method also starts by assigning the leaf nodes to the bottom level, and then performs a BFS to define the level of a non-bottom node as its shortest distance from a bottom one [20]. In contrast, the hierarchical score maximization (HSM) algorithm presented here works to globally optimize the hierarchy of a directed network. It defines a hierarchy score (HS) to quantify the degree of hierarchy in a network. The hierarchical score captures the global hierarchical property of a network. To infer the hierarchy, HSM optimizes the hierarchical structure so that the maximum HS is achieved. Thus, the hierarchy inferred by HSM represents the globally optimized structure. The VS algorithm identifies strongly connected components and collapses them to convert the network into a directed acyclic graph, and applies the leaf removal algorithm on the graph and on its transpose. Results are then combined to infer a global solution of hierarchical levels [21]. This method is not designed to maximize the downward information flow and thus the resulting networks have smaller hierarchy score compared to those from the HSM algorithm. [[Ref1.3.1]]

Compared with the previous methods, the HSM algorithm takes into account the potential hierarchical ambiguity underlying a network. It provides a probabilistic representation of the

hierarchy for a network that can more precisely reflect the underlying hierarchical structure. For all the nodes, we know the certainty of them being assigned to a hierarchical level, which is informative and is useful for us to interpret their roles in the hierarchical network. A global optimization method has been proposed in [27], which applied a simulated annealing algorithm to minimize the number of "backward" links going from lower to higher hierarchical levels. In contrast, we define a hierarchical score that quantifies the degree of hierarchy and infer the hierarchical structure of a network by maximize this score. [[Ref1.3.2]]

Moreover, the HSM algorithm's corrected hierarchical score (CHS) is comparable between different networks. As shown in Table 1, this enables comparisons in the degree of hierarchy between different biological networks such as social networks, file sharing networks, ecological networks, and neural networks. Practically, this allows for the exploration of the common rules shared by different networks and reveals the differences between them [31]. For example, we find that the protein phosphorylation interactions mediated by kinases are much more hierarchical than the transcriptional regulatory interactions mediated by TFs.

Hierarchy versus asymmetry for directed network

Dyadic reciprocity and Krackhardt hierarchy score are often used to quantify the extent of asymmetry in directed networks [32]. The former is defined as the proportion of node pairs that are reachable from either direction, while the latter is the fraction of node pairs that are reachable from only one direction. We note that Krackhardt hierarchy score, though termed as a "hierarchy" score, is distinct from the hierarchy score (HS) described here. The asymmetry measured by reciprocity or Krackhardt score quantifies how possible two nodes are "mutual reachable" in a directed network. Another metric called global reaching centrality (GRC) was defined to measure hierarchy as heterogeneous distribution of the local reaching centrality (the proportion of all nodes that can be reached from a node) of all nodes in a directed network[33]. These metrics does not imply any information on orientation. In contrast, by hierarchy here we mean a top-to-bottom orientation for nodes at different levels.

Why do we need to introduce the "orientation/hierarchy" attribute for a directed network? Because in many networks the nodes are by nature associated with certain "spatial" or "temporal" attributes. For example, the protein nodes in a biology network may localize in different cellular components, e.g. the membrane, cytoplasm, or nucleus etc; meanwhile, external signals are often transduced following a specific direction from membrane to nucleus. This confers a global "hierarchy" attribute to the network that cannot be captured by "asymmetry" attributes (e.g. reciprocity and Krackhardt score). On the other hand, since the "hierarchy" originates from certain attributes of nodes, we would expect to observe the correlation of hierarchy with node features. In other words, the inferred hierarchical structure should recapitulate the attribute difference of nodes at different levels. For instance, as shown in Fig. 4, we find that the higher-level TFs in the yeast regulome are more likely to be essential and more conserved.

The hierarchy score is also different the three-dimensional "morphospace" proposed recently by Corominas-Murtra, which defines three hierarchical features: treeness, feedforwardness and orderability [34]. To define them, nodes with zero in-degrees and out-degrees are regarded as the source and the sink of a network, respectively, and then the paths between them are characterized.

Temporal versus spatial organization of hierarchy

~~All of the previous hierarchical network studies have focused on the regulome. In this study, we construct the hierarchy of the yeast phosphorylome and correlate it with a number of kinase properties, including essential rate, conservation and so on.~~

We observe interesting results when we apply the HSM algorithm to the yeast regulome and phosphorylome. In the yeast regulome, we find that higher-level TFs are more likely to be essential and are more conserved during evolution. Particularly, none of the 62 bottom-level TFs are essential, compared to an average essentiality rate of 19% in yeast. In the yeast phosphorylome, however, this is not the case and instead we observe significant differences in cellular localization for kinases at different levels. For instance, 21% of the middle-level and 18% of the bottom-level kinases are detectable in bud/bud-neck, whereas in the 38 top-level kinases, none are identified in bud/bud-neck.

Biologically, the hierarchy of regulatory networks (regulome and phosphorylome) may arise from the temporal and/or spatial organization of regulators. In response to stimulation or in a biological process (e.g. cell cycle regulation), early-activated regulators (e.g. TFs or kinases) regulate the expression/activation of later regulators, which in turn regulate even later ones, forming a hierarchical structure. Similarly, the cellular localization of regulators can also contribute to the hierarchical organization of a regulatory network. For example, during signal transduction the extracellular signal is typically transferred from a membrane-localized kinase to a cytoplasmic kinase and onward to a nuclear kinase [35]. Since in general TFs function in nucleus by regulating gene expression, their hierarchy is mainly organized via temporal activation of TFs. However, in phosphorylomes the hierarchical organization of kinases can be determined by both temporal regulation and spatial localization. The differential correlation pattern of protein features with hierarchy between regulomes and phosphorylomes may reflect such a difference.

In summary, the HSM algorithm provides a useful tool to investigate the hierarchy of directed networks. It can be used independently or in conjunction with other hierarchy inference methods. With more and various regulatory interaction data coming out, we expect a wide application of these methods in biological network studies. [[Ref1.1.3]]

Methods

Construction of network hierarchy

A hierarchical network is a directed network for which all nodes are assigned to a unique hierarchical level from 1 to L, where L is the total number of levels (L≥2). Generally, a hierarchical network contains three types of edges according to their directionality: a downward edge (pointing from a higher level node to a lower level node), an upward edge (pointing from a higher level node to a lower level node), and a horizontal edge (pointing from a node to another node in the same level). To infer the hierarchy of a directed network, we developed a hierarchical score maximization algorithm described as follows.

First, given a directed network with assigned hierarchical structure, we define a metric called hierarchy score as:

$HS = \frac{N_d + N_h}{N_u + N_h}$, where N_d, N_u, and N_h are the number of downward edges, upward edges and horizontal edges, respectively. The metric essentially measures the ratio of N_d to N_u balanced by N_h. It takes a value from 0 to +∞, with a higher HS indicating more downward edges relative to upward edges in a network. Specifically, when $N_u=N_h=0$, the network will have a HS of +∞. (Ref1.1)

Second, for a directed network we employ a simulated annealing procedure [36] to infer its hierarchical structure by arranging nodes into L levels (L is a pre-defined parameter). This procedure is as follows:

(1) We initiate from randomly assigning each node to a level, calculate the corresponding HS score hs_0, and setting the initial energy as $E_0 = -hs_0$.
(2) We adjust the hierarchy iteratively to optimize the hierarchical structure. Specifically, at iteration i, we randomly select a node, adjust the hierarchy by randomly placing it into another level and recalculate the hierarchical score and energy ($hs_i = -E_i$) of the resulting new hierarchy. We compute the energy change $\Delta E = E_i - E_{i-1}$; if $\Delta E<0$, we accept the hierarchy adjustment; otherwise we accept the adjustment with a probability $P=\exp(-\Delta E/CT)$, where C is a constant and T is temperature that are used to tune the probability P.
(3) We repeat this procedure p times until E is minimized (i.e. HS is maximized). In practice, we gradually lower the temperature T at each step to adjust the sensitivity of annealing. This procedure results in an optimized hierarchical network with maximized HS score.

Third, we perform the above-described simulated annealing algorithm k=1000 times to obtain 1000 inferred hierarchical networks. We do this because in many cases the optimum hierarchy is not unique. For example, some nodes are topologically identical in a directed network, and changing their level assignment coordinately will not change the overall hierarchical score. Based on these 1000 inferred hierarchical networks, we calculate the probability that each node is assigned to each level, which results in a probability matrix for each node as seen in Fig. 1C. This matrix can be regarded as a probabilistic hierarchical network, which is more informative and more precisely describes the hierarchical structure of a directed network than methods that omit this procedure.

Fourth, we provide a most likely hierarchical network based on the probabilistic hierarchical matrix. Specifically, we assign each node to the level for which the prior step assigns it the highest probability. It should be noted that the confidence of the assignment might vary from node to node, depending on the value of the maximum probability. Typically, however, most of the nodes have high certainty in terms of the level assignment (e.g. the probability in the assigned level is >60%).

To determine an appropriate p (the number of steps in each simulated annealing procedure) and k (the number of each simulated annealing runs), we plot the hierarchy score against p and k, respectively. For a network with more nodes and edges, a larger p should be used as can be determined based on the HS vs. p plot. When a suitable p is used, the resulting HS should be stable against k when k is >100. [[Ref1.1.4]]

In practice, the HSM method can be used conjunction with other hierarchy inference methods. For example, one may start from the hierarchical structure inferred by the VS algorithm, and use the simulated annealing procedure method to further optimize the hierarchy score. Namely, instead of randomly selecting nodes during the simulated annealing optimization, we can focus on adjusting the levels of ambiguous nodes from VS output to improve the efficiency. Such a strategy will combine the advantages of the two hierarchy inference approaches. [[Ref1.1.3]]

Determination of the number of hierarchical levels

The HSM algorithm requires a pre-defined L, the number of hierarchical levels. L can be determined based on the prior knowledge about the directed network of interest. If no prior knowledge is available, we can specify different L values (e.g. L=2, 3, …, 8) and choose a proper L by comparing the resulting hierarchical networks. However, the HS score is not directly comparable for hierarchical networks with different number of levels, because networks with larger L tend to have higher HSs. We thereby define a corrected hierarchical score (CHS) as the following:

$$CHS = \frac{O(N_d)/E(N_d) + O(N_h)/E(N_h)}{O(N_u)/E(N_u) + O(N_h)/E(N_h)},$$

where $O(N_d)$, $O(N_u)$ and $O(N_h)$ are the observed number of downward, upward and horizontal edges, respectively; $E(N_d)$, $E(N_u)$ and $E(N_h)$ are the expected number of downward, upward and horizontal edges, respectively. $E(N_d)$, $E(N_u)$ and $E(N_h)$ are calculated as:

$$E(N_d) = \sum_{i>j} S_i S_j ;$$

$$E(N_u) = \sum_{i<j} S_i S_j ;$$

$$E(N_h) = \sum_{i=j} S_i S_j ,$$

where S_i and S_j are the number of nodes in level *i* and level *j*, respectively. The CHS is directly comparable between hierarchical networks with different L values, and can also be used to compare the degree of hierarchy between different directed networks. The CHS takes a value from 1 for random network without a hierarchical structure to ∞ for a network with a perfect hierarchy (e.g. a tree as in Fig. 2).

To determine the number of hierarchical level L for a network, one can employ the HSM algorithm across a range of L values and choose the L for which the HSM algorithm yields the highest CHS. In some cases, the CHS will keep increasing with the increase of L, because there is more freedom to optimize the hierarchy with larger L values. In this situation, one can plot the CHS against L values, and choose the L at which no significant CHS improvement is achieved. In addition, other information is also important to determine the L for a directed network. For example, it is reasonable to require L to be no larger than the diameter of the network, namely, the greatest distance between any pair of connected nodes.

Calculation of probabilistic hierarchical score

To more accurately measure the hierarchical structure of a probabilistic hierarchical network, we define a new metric called the probabilistic hierarchical score (PHS). For an edge $i \to j$ in a network with L levels, the probability of this edge being downward is $\sum_{L_i>L_j} P(L_i,i)P(L_j,j)$, where $P(L_i,i)$ and $P(L_j,j)$ are the probability of the node i and j in level L_i and L_j, respectively. Similarly, the probability of $i \to j$ to be upward is $\sum_{L_i<L_j} P(L_i,i)P(L_j,j)$; and the probability of $i \to j$ to be horizontal is $\sum_{L_i=L_j} P(L_i,i)P(L_j,j)$. Thus after taking into account all edges in the network, we define PHS as the following:

$$PHS = \frac{\sum_{(i\to j)\in\{e\}} \sum_{L_i>L_j} P(L_i,i)P(L_j,j) + \sum_{(i\to j)\in\{e\}} \sum_{L_i=L_j} P(L_i,i)P(L_j,j)}{\sum_{(i\to j)\in\{e\}} \sum_{L_i<L_j} P(L_i,i)P(L_j,j) + \sum_{(i\to j)\in\{e\}} \sum_{L_i=L_j} P(L_i,i)P(L_j,j)}.$$

The level is indexed in an increasing order from bottom to top. Namely, level *i* is higher than level *j* in the hierarchy, if $i>j$.

Estimation of the hierarchy significance for a directed network

Given a directed network, the HSM algorithm infers its optimum hierarchical structure by maximizing the HS score. Although the resulting HS score can measure the degree of hierarchy of a network, it does not tell us whether a directed network has a significantly hierarchical structure. To address this issue, we compare a directed network with random networks to evaluate its hierarchical significance. Here we use the Erdos-Renyi random graph model as the null model. In a network, each pair of nodes has an equal chance to be connected by an edge [37]. We generate 1000 Erdos-Renyi random networks with the same number of nodes and edges, and calculate their HSs using the HSM algorithm. The P-values of hierarchy

for the network of interest is then computed as the fraction of random networks with a HS equal to or greater than the interested network. Alternatively, assuming a Gaussian distribution of the HSs of the random networks, we calculate the Z-score for the interested network: $Z=(HS-\mu)/\sigma$, where μ and σ are the mean and standard deviation of the HS scores of those random networks; the P-value is calculated by referring to a standard normal distribution. We note that the significance estimation depends on the selection of the null model. To generate the random networks, other null models can be used and certain constraints can be applied as required.

Calculation of dyadic reciprocity and Krackhardt hierarchy score

Traditionally, 1-dyadic reciprocity and Krackhardt hierarchy score are often used to quantify the extent of asymmetry in directed network [32]. The dyadic reciprocity is defined as the proportion of node pairs in a directed network that are symmetric (i.e. reachable from either direction). Krackhardt hierarchy score is the fraction of node pairs in the directed network that are reachable from one direction. These two metrics measure the degree of asymmetry of a directed network, which is different from the hierarchy we introduce in this study. Our hierarchy by nature implies a top-to-bottom orientation, whereas the "asymmetry" is non-directional. We use the R package "sna" to calculate dyadic reciprocity and Krackhardt hierarchy score. The global reaching centrality of networks is calculated using the method introduced by Mones et al [33].

Directed networks used in this study

In this study, we examine eight directed networks, including five biological networks, one ecological network (food web network), one social network (political blogs network) and one computer network (P2P file sharing network). The five biological networks are the yeast regulation network, the human regulation network, the yeast phosphorylation network, the human phosphorylation network and the worm neural network.

The yeast regulome was downloaded from Jothi et al [21], in which most of the TF-gene interactions were identified by ChIP-chip experiments [9, 10], and the remaining were collected based on other biochemical studies [38-41]. The human regulome is constructed based on ChIP-seq data from the ENCODE project [42], based on which the target genes of more than 120 TFs are determined by a probabilistic model [43]. For TFs with multiple ChIP-seq datasets, the target genes represent a union of targets in all of the available datasets. The kinase-substrate interactions in the yeast phosphorylome are collected from protein chip experiment by Ptacek et al [11] and the phosphorylation site data collected by Freschi et al. from several large-scale studies [44]. The human phosphorylome is available from the PhosphoNetworks database (http://phosphonetworks.org), which is based on experimental determined kinase-substrate relationships [37]. In our hierarchical study, we include only TF-TF interactions in the two regulomes and kinase-kinase interactions in the two phosphorylomes.

The worm neural network contains the interaction of one neuron to another via synaptic or gap junctions in worm [45]. The food web network is from Ulanowicz et al, which contains the carbon exchange from one species to another occurring during the wet season in the cypress wetlands of south Florida [46]. The Political blogs network contains hyperlinks between weblogs on US politics being recorded in 2005 [3]. The P2P file-sharing network is one of a series of Gnutella network created in 2002, in which nodes represent host computers in the Gnutella computer network and edges represent connections between the hosts [47].

Properties of yeast genes and proteins

The list of yeast essential genes was determined by a yeast gene deletion project and was downloaded from the Saccharomyces genome database (SGD) [48]. The Ka/Ks ratios of ortholog genes between S. *cerevisiae* and S. *pombe* orthologs were from Wall et al [49]. The physical and genetic interactions of yeast genes were also downloaded from the SGD database [50, 51]. Specifically, the protein-protein interactions between yeast kinases were obtained from Breitkreutz et al [12]. The mRNA abundance and mRNA half-life data were obtained from previous studies (Holstege et al, 1998; Wang et al, 2002). The protein half-life data came from Belle et al [52]. The protein abundance and protein noise data were available from Newman et al [53]. To determine the protein noise, the single cell expression level of a protein was measured in a population of yeast cells and then the ratio of the standard deviation to its mean abundance was calculated. For a protein, the noise is represented as the difference between its noise value and the median over all proteins, named as deviation from median (DM). Budding or budding neck localization of yeast kinases was obtained from Huh et al [54]. The cellular component associated with yeast kinases was annotated by SGD, which are manually curated based on previous publications.

Enrichment of interactions between different levels

To examine whether TFs/kinases are more likely to physically/genetically interact within the same level or between two levels (Fig. 6A), we calculated the enrichment of interactions in all pairs of levels: TT, TM, TB, MM, MB and BB. Using physical interactions between TFs as the example, the significance of enrichment or depletion is calculated as follows. First, given a physical interaction network with n nodes and e edges, we compute the probability for a pair of randomly selected genes to interact: p=e/[n(n-1)/2]. Second, we assume that the number of TF-TF interactions (denoted as i) within a level or between two different levels follows a binomial distribution: $\Pr(x = i) = f(i;b,p) = C_b^i p^i (1-p)^{b-i}$, where b is the number of all possible TF-TF pairs. Considering self-interactions, b=m(m+1)/2 for intra-level interactions with m TFs (i.e., TT, MM or BB), and $b=m_1m_2$ for interactions between two levels with m_1 and m_2 TFs, respectively (i.e., TM, TB and MB). Finally, the p-values are calculated as $P(x \geq i)$ for enrichment (i.e., the probability of observing an equal or greater number of interactions) and $P(x \leq i)$ for depletion (i.e., the probability of observing an equal or smaller number of interactions) of physical interactions between these TFs.

To estimate whether two kinases share a significantly large number of physical partners, genetic partners or substrates (Fig. 6B), we examine their degree of overlap and calculate its significance using the Fisher exact test (i.e. hyper-geometric test).

Gene ontology analysis

We used the DAVID Gene Ontology (GO) annotation tool [55] to investigate the functional enrichment of kinases in the three levels of our hierarchical network for phosphorylome (Fig. 3B). The whole list of the 94 kinases in the network is used as the background for enrichment analysis. A similar analysis is also used to study the functional enrichment of substrates specific to kinases from each of the three levels. In this case, we use the whole yeast gene list as the background.

Authors' contributions

CC conceived of the study, participated in its design and coordination, performed most of the analysis, and drafted the manuscript. MG conceived of the study, participated in its design and coordination, and helped to draft the manuscript. EA participated in the data analysis, and helped to draft the manuscript. KY, MU and DW participated in its design, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge support from the NIH and from the AL Williams Professorship funds. This work was also supported by the start-up funding package provided to C.C. by the Geisel School of Medicine at Dartmouth College.

References

1. Barabasi AL, Albert L, Jeong H, Bianconl G: **Power-law distribution of the world wide web.** *Science* 2000, **287:**2115.
2. Albert R, Jeong H, Barabasi AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382.
3. Adamic LA, Glance N: **The political blogosphere and the 2004 US Election.** In *WWW-2005 Workshop on the Weblogging Ecosystem* 2005
4. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654.
5. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
6. Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C: **Systematic mapping of genetic interaction networks.** *Annu Rev Genet* 2009, **43:**601-625.
7. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nat Biotechnol* 2000, **18:**1257-1261.
8. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, et al: **Global mapping of the yeast genetic interaction network.** *Science* 2004, **303:**808-813.
9. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
10. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431:**99-104.
11. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, et al: **Global analysis of protein phosphorylation in yeast.** *Nature* 2005, **438:**679-684.
12. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G, et al: **A global protein kinase and phosphatase interaction network in yeast.** *Science* 2010, **328:**1043-1046.
13. Bulyk ML: **Discovering DNA regulatory elements with bacteria.** *Nat Biotechnol* 2005, **23:**942-944.
14. Ouwerkerk PB, Meijer AH: **Yeast one-hybrid screening for DNA-protein interactions.** *Curr Protoc Mol Biol* 2001, **Chapter 12:**Unit 12 12.
15. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
16. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316:**1497-1502.
17. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan KK, Dong X, Djebali S, Ruan Y, et al: **Understanding transcriptional regulation by integrative analysis of transcription factor binding data.** *Genome Res* 2012, **22:**1658-1667.
18. Shou W, Verma R, Annan RS, Huddleston MJ, Chen SL, Carr SA, Deshaies RJ: **Mapping phosphorylation sites in proteins by mass spectrometry.** *Methods Enzymol* 2002, **351:**279-296.
19. Bhardwaj N, Kim PM, Gerstein MB: **Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators.** *Sci Signal* 2010, **3:**ra79.
20. Yu H, Gerstein M: **Genomic analysis of the hierarchical structure of regulatory networks.** *Proc Natl Acad Sci U S A* 2006, **103:**14724-14731.
21. Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L, Babu MM: **Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture.** *Mol Syst Biol* 2009, **5:**294.

22. Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, et al: **Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.** *PLoS Comput Biol* 2011, **7:**e1002190.
23. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.** *Science* 2010, **330:**1775-1787.
24. Ma HW, Buer J, Zeng AP: **Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach.** *BMC Bioinformatics* 2004, **5:**199.
25. Ma HW, Kumar B, Ditges U, Gunzer F, Buer J, Zeng AP: **An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs.** *Nucleic Acids Res* 2004, **32:**6643-6649.
26. Hartsperger ML, Strache R, Stumpflen V: **HiNO: an approach for inferring hierarchical organization from regulatory networks.** *PLoS One* 2010, **5:**e13698.
27. Ispolatov I, Maslov S: **Detection of the dominant direction of information flow and feedback links in densely interconnected regulatory networks.** *BMC Bioinformatics* 2008, **9:**424.
28. Sakata S, Yamamori T: **Topological relationships between brain and social networks.** *Neural Netw* 2007, **20:**12-21.
29. Mok J, Zhu X, Snyder M: **Dissecting phosphorylation networks: lessons learned from yeast.** *Expert Rev Proteomics* 2011, **8:**775-786.
30. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al: **The genetic landscape of a cell.** *Science* 2010, **327:**425-431.
31. Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M: **Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.** *Proc Natl Acad Sci U S A* 2010, **107:**9186-9191.
32. Krackhardt D: **Graph Theoretical Dimensions of Informal Organizations.** In *Computational Organization Theory.* In K. M. Carley and M. J. Prietula edition: Hillsdale, NJ: Lawrence Erlbaum and Associates; 1994: 89-111
33. Mones E, Vicsek L, Vicsek T: **Hierarchy measure for complex networks.** *PLoS One* 2012, **7:**e33799.
34. Corominas-Murtra B, Goni J, Sole RV, Rodriguez-Caso C: **On the origins of hierarchy in complex networks.** *Proc Natl Acad Sci U S A* 2013, **110:**13316-13321.
35. Roberts PJ, Der CJ: **Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer.** *Oncogene* 2007, **26:**3291-3310.
36. Kirkpatrick S, Gelatt CD, Jr., Vecchi MP: **Optimization by simulated annealing.** *Science* 1983, **220:**671-680.
37. Newman RH, Hu J, Rho HS, Xie Z, Woodard C, Neiswinger J, Cooper C, Shirley M, Clark HM, Hu S, et al: **Construction of human activity-based phosphorylation networks.** *Mol Syst Biol* 2013, **9:**655.
38. Svetlov VV, Cooper TG: **Review: compilation and characteristics of dedicated transcription factors in Saccharomyces cerevisiae.** *Yeast* 1995, **11:**1439-1484.
39. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M: **Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae.** *Genes Dev* 2002, **16:**3017-3033.
40. Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M: **Target hub proteins serve as master regulators of development in yeast.** *Genes Dev* 2006, **20:**435-448.
41. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M: **Divergence of transcription factor binding sites across related yeast species.** *Science* 2007, **317:**815-819.

42. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
43. Cheng C, Min R, Gerstein M: **TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.** *Bioinformatics* 2011, **27:**3221-3227.
44. Freschi L, Courcelles M, Thibault P, Michnick SW, Landry CR: **Phosphorylation network rewiring by gene duplication.** *Mol Syst Biol* 2011, **7:**504.
45. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393:**440-442.
46. Ulanowicz RE, Heymans JJ, Egnotovich MS: **Network Analysis of Trophic Dynamics in South Florida Ecosystems, FY 99: The Graminoid Ecosystem.**; 2000.
47. Leskovec J, Kleinberg J, Faloutsos C: **Graph Evolution: Densification and Shrinking Diameters.** In *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, vol. 1; 2007.
48. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al: **Saccharomyces Genome Database: the genomics resource of budding yeast.** *Nucleic Acids Res* 2012, **40:**D700-705.
49. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proc Natl Acad Sci U S A* 2005, **102:**5483-5488.
50. Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D: **Genetic and physical maps of Saccharomyces cerevisiae.** *Nature* 1997, **387:**67-73.
51. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, et al: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.** *Science* 1999, **285:**901-906.
52. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK: **Quantification of protein half-lives in the budding yeast proteome.** *Proc Natl Acad Sci U S A* 2006, **103:**13004-13009.
53. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: **Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise.** *Nature* 2006, **441:**840-846.
54. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
55. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.

Table 1. Hierarchical scores of eight directed networks.

	#nodes	#edges	#levels	1-DR	KHS	GRC	HS	CHS	PHS	Reference
Worm neural	297	2359	L=4	0.184	0.188	0.133	2.835	2.384	2.707	Watts et al.
Political blogs	1224	19087	L=3	0.487	0.517	0.130	3.011	3.177	2.972	Acamic et al.
Yeast TF	149	580	L=4	0.559	0.611	0.381	4.675	3.669	4.330	Harbison et al.
Human TF	772	518	L=4	0.631	0.718	0.336	7.097	5.606	5.846	Gerstein et al.
P2P file sharing	5301	20777	L=4	0.486	0.772	0.628	4.346	5.676	2.401	Ripeanu et al.
Foodweb	83	612	L=3	0.259	0.261	0.582	5.795	6.407	5.786	Ulanowicz et al.
Human Kinase	373	2171	L=4	0.492	0.758	0.020	14.087	13.346	12.874	Newman et al
Yeast Kinase	84	200	L=4	0.645	0.775	0.447	17.455	13.982	11.777	Ptacek et al

1-DR: 1-dyadic reciprocity; KHS: Krackhardt hierarchy score; GRC: global reaching centrality; HS: hierarchy score; CHS: corrected hierarchy score; PHS: probabilistic hierarchy score.

Legend of Figures:

Figure 1: The schematic diagram of the hierarchy score maximization algorithm. In hierarchical networks, the downward, upward and horizontal edges are shown in red, blue, and black colors, respectively. (A) The definition of hierarchy score. (B) A simulated annealing algorithm for inferring the hierarchical structure by maximizing the hierarchy score. (C) The procedure to calculate the probability of nodes in different hierarchy levels. Simulated annealing procedure is performed for k runs and in each run a hierarchical structure is inferred by maximizing the hierarchy score. The frequency of each node in these k hierarchical networks is calculated to obtain a probabilistic hierarchical network. Discretized hierarchical network is obtained by assigning nodes to the level with highest frequency. (Ref1.1.4)

Figure 2: Application of the hierarchy score maximization algorithm to a military command network. (A) A military command network with 19 nodes at 5 hierarchy levels. (B) The probability matrix inferred by the HSM algorithm with the number of levels specified as L=2, 3, …, 8. Each element in the matrix gives the probability of a node being assigned to a level. The HSM algorithm correctly identifies the network hierarchy when L=5 is specified. (C) The distribution of hierarchy scores when a certain number of edges in the original network are perturbed. HS: hierarchy score; CHS: corrected hierarchy score; PHS: probabilistic hierarchy score (see "Methods" for details)

Figure 3: Application of the HSM algorithm to the yeast regulome (A), phosphorylome (B) and a random network (C).

Figure 4: Application of HSM algorithm to the yeast regulome. (A) The corrected hierarchy scores for hierarchical networks as inferred by HSM, BFS and VS methods. (B) The number of downward, upward and horizontal edges in hierarchical networks inferred by the three methods. (C) The correlation of TF properties with hierarchy. T, M and B represent top-, middle-, and bottom- level, respectively.

Figure 5: Application of the HSM algorithm to the yeast phosphorylome. (A) The localization of kinases at different levels in the cytoplasm and nucleus. (B) The correlation of kinase properties with hierarchy. T, M and B represent top-, middle-, and bottom- level, respectively.

Figure 6: Collaboration of kinases at different hierarchy levels. (A) The enrichment of physical and genetic interactions of kinases within a level (TT, MM and BB) and between two levels (TM, TB, and MB). (B) The enrichment of kinase pairs with significantly overlapping physical or genetic interaction partners or phosphorylation substrates. (C) The enrichment of positive and negative genetic interactions of kinases. Enrichment and depletion of interactions ($P<0.05$) are marked as red and green "*" respectively.

Figure 7: Properties of the inferred hierarchical structure for yeast phosphorylome generated by HSM algorithm. (A) The distribution of feed-forward loop (FFL) motifs in the hierarchical network. In a FFL motif, a kinase X phosphorylates another kinase Y and both target a common substrate Z. Depending on the location of X and Y in the hierarchical structure, the X-->Y interaction can be categorized into 9 combinations. Downward interactions (TM, TB and MB), upward interactions (MT, BT and BM), and horizontal interactions (TT, MM and

BB) are shown with red, blue and gray bars, respectively. (B) The Venn diagram of substrates of kinases at different levels.