Automatic discovery of case-specific relevant subgraphs in large scale signaling networks using random walks

Aristotelis Kittas[1], Carito Guziolowski[2], Niels Grabe[3]

[1]Hamamatsu Tissue Imaging and Analysis Center, Bioquant, Department of Medical Oncology, National Center for Tumour Diseases, Heidelberg University, Im Neuenheimer Feld 267 (BQ 0010), D-69120, Heidelberg, Germany.

[2]École Centrale de Nantes, IRCCyN UMR CNRS 6597 (Institut de Recherche en Communications et Cybernétique de Nantes) 1 rue de la Noë - B.P. 92101 - 44321 Nantes Cedex 3, France.

Abstract

We present a method to discover signaling pathways, quantify the relationship of preselected source/target nodes, and extract relevant subgraphs in large scale biological networks. This is demonstrated over the hepatocyte growth factor (HGF) stimulated cell migration and proliferation in a keratinocyte-fibroblast co-culture. The algorithm (MCWalk) is implemented with random walks using Monte Carlo simulations. We extract a master network by overlaying case specific microarray data from the NCI Pathway Interaction Database (PID) using a fully automatic pipeline without any manual network construction, and uncover the association of HGF receptor c-Met nodes, differentially expressed (DE) protein nodes and cellular states. We show that the network has a scale-free structure and identify key regulator nodes based on their random walk traversal frequency. This property is shown to be very weakly correlated to node degree, contrary to what is expected from similar centrality measures. The differences with standard methods, such as shortest-path, commonly used in the analysis of such networks are discussed and compared with this approach, highlighting important pathways which are exclusively obtained with our random walks algorithm.

Introduction

The cell receives, responds and processes information via a variety of signaling pathways. The components of different pathways interact with each other, giving rise to signaling networks. Such interactions include mechanisms like regulation of protein-protein interactions, phosphorylation, regulation of enzyme activity, production of secondary messengers etc. Systems biology teaches us not only that phenotypes are influenced by many important genetic and environmental factors, but also that we need to understand the extremely complex interactions

[1]e-mail: akittas@gmail.com

between these components [1]. The representation of biological systems as large scale networks has thus become increasingly popular, in order to evaluate biological properties based on the interaction of signaling pathways. Relatively few methods have been proposed so far for analyzing the structure of a given signaling (or any interaction) network [2]. Structural analysis is particularly useful in large networks, where a simple visual inspection is not possible and the construction of quantitative models is practically infeasible due to the large amount of parameters required.

The method of random walks has been well-established for structural analysis of networks, as it can fully account for local as well as global topological structure within the network and is useful for identifying most important/central nodes [3–5]. Random walks arise in many models of mathematics and physics and their properties have been studied in systems of various geometries [6]. They are closely related to centrality in networks, which is widely used for measuring the relative importance of nodes within a graph [20]. Noh and Rieger [26] introduced random walk centrality in undirected graphs. This quantifies how central a node *u* is located regarding its potential to receive information randomly diffusing over the network. It is a measure of the "speed" with which randomly walking messages reach a vertex from elsewhere in the network, a sort of random-walk version of closeness centrality.

Newman's random walk betweenness [27] is a similar notion, which quantifies the number of times that a random walk starting at *s* and ending at *t* passes through a node *u* along the way, averaged over all *s* and *t*. This is more closely related to Freeman's betweenness centrality; one end representing information that has no idea of where it is going and the other information that knows precisely where it is going (i.e. traversing the network on shortest paths). Estrada's communicability betweenness [28] combines these two measures, allowing information to pass through all possible routes, but introducing a scaling so that longer walks carry less importance.

A random walk is a finite Markov chain and in fact all Markov chains can be viewed as random walks on a directed graph. In physics random walks in graphs have been widely studied, especially in the diffusion reaction scheme (e.g. [7–12]). These processes have many applications in communications and social networks (e.g. information propagation and rumor spreading), however few studies [5,13] have applied them in the analysis of the structure and function of biological, and in particular, signaling networks.

A notable exception is the field of subgraph extraction, which can be used to predict a meaningful pathway given a biological network (e.g. protein–protein interaction or metabolic network) and a set of query items (e.g. genes, proteins and compounds) defining seed nodes in the network [14]. One is then interested in obtaining a subgraph that best captures the relationships between *k* given nodes of

interest (or seed nodes) in a graph. This is a powerful technique, which can be used to predict pathways from biological networks and a set of query items (e.g genes, proteins, compounds etc). The most straightforward method to obtain such a network is to connect all nodes of interest using the shortest path. Dijkstra's algorithm [15] is widely used to obtain such paths in a network.

This is often a reasonable simplification in social and communications networks, where information often propagates along shortest paths. However, the same cannot be said about biological networks, where the length of the path is often of no biological significance and using such algorithms is an oversimplification which often results in omitting important pathways. Therefore, random walk methods have been developed to improve the completeness and relevance of the extracted paths in the network. Such methods often use clustering algorithms like Markov Clustering (MCL) [16] and have recently been applied to split biochemical networks into coherent subnets [17]. A clustering method somewhat similar to MCL has also been applied to metabolic networks [18]. The NetWalk algorithm [5] has been developed to calculate the distribution of edge flux values associated with each interaction in the network, which reflects the relevance of interactions based on the experimental data.

The discovery of all possible (non-simple) paths between any set of nodes is a particularly difficult case and known to be an NP-complete problem in cyclic graphs [34]. Various methods have been proposed to extend the completeness of the shortest path approach, e.g. by discovering the k-Shortest paths (i.e. the second, third, etc... path), and such algorithms have been recently applied to infer regulatory pathways in a gene network [35]. Dupont et al [19] implemented a generic algorithm (k-walks), to build a subgraph connecting seed nodes, which contains the most relevant edges and the nodes induced by those edges. The relevance of an edge is measured as the expected number of times it is visited along random walks connecting seed nodes, which reflect both the topology of the network and the edge weights.

The k-walks algorithm [13,19] simulates random walks on the network using Markov chains, computing the set of edges/nodes most likely to be used while randomly walking between seed nodes. This method finds a "relevant subgraph", which should be as small as possible while capturing most of the information between *k* nodes of interest (seed nodes). They follow from an interpretation of the graph as a Markov chain characterized by a transition probability matrix ***P***. The probability of transition from node *i* to node *j* is then given by: $P_{i,j} = \frac{w_{i,j}}{\sum_j w_{i,j}}$, where $w_{i,j}$ is the weight assigned to edge $i \rightarrow j$.

To formalize this idea, it's required that a random walker which starts in any node of interest should be able to reach at least another node of interest, in order to explain the relationship between all nodes (e. g. genes in a regulatory network). A subgraph is then obtained by keeping only those edges above a minimal relevance threshold. This method, combined with shortest-path methods has been shown to be fairly accurate in metabolic pathway prediction [13]. While this approach is very valuable in analyzing gene regulation and metabolic networks, it might have limitations in signaling networks where terminal nodes of interest may exist.

In our case, we frequently map unidirectional relationships (e.g. a protein affecting a cellular process, or a transcription factor regulating the expression of a gene) between nodes in pathway interaction databases, such as the PID. In specific, we investigate the relationship between growth factors, differentially expressed genes and cell states, such as proliferation and migration. It is thus possible (and a frequent case) that in a growth factor signaling pathway a molecule is linked to another in a unidirectional relationship (e.g. a GF to a DE gene or a cellular state). In signaling networks, many nodes of interest (e.g. cellular states) are terminal nodes, i.e. they don't have any successors, or there isn't any node of interest that is reachable from them. There are also cases when we want to map causal relationships between specific sets of nodes, which can act exclusively as source or target nodes. In these scenarios we are unable to use a method like kWalks, which requires that a directed path must exist from each node of interest to at least one other node of interest and that each of these nodes should have degree $k \geq 2$. This is because the initial state of the Markov Chain cannot be an absorbing state, and other states (including the initial state) form the set of transient states, from which there is a strictly positive probability to leave.

Thus, we developed an algorithm based on random walks using Monte Carlo simulations (from here on referred to as "MCWalk"), that links sets of predefined input/output nodes, quantifies their relation and extracts relevant functional sub-networks. The MCWalk algorithm allows the existence of terminal nodes and in addition, the extraction of causal relationships between the source/target nodes based on the structure of the network, obtaining a quantitative estimate of the correlation of these nodes based on frequency of traversal. We demonstrate this on hepatocyte growth factor (HGF) stimulated cell migration and proliferation in a keratinocyte-fibroblast co-culture time series data from Busch et al. [29]. Hepatocyte growth factor (HGF) and its receptor, Met, regulate a number of biological functions in epithelial and nonepithelial cells, such as survival, motility, proliferation, and tubular morphogenesis [30]. We obtain a family of functionally relevant signaling networks based on the traversal of a random signal between the specified source/target nodes for HGF signaling. This method can thus be applied in interpreting any experiment results, where a causal relation needs to be established

between sets of input/output nodes and constructing case-specific networks that model the underlying biological mechanisms.

Methods

Pipeline description

An automatic pipeline introduced in [31] is used to obtain computable networks from the Pathways Interaction Database (PID) [32], which are readable in Cytoscape [33], and link the signaling events of specific growth factors to the gene response observed in a particular gene expression experiment. The pipeline requires the following information:

- *Network:* Complete PID NCI-Curated xml download.
- *Seed nodes:*
    - HGF receptor c-Met nodes
    - Differentially expressed (DE) protein nodes: these were molecular nodes of type protein that met two conditions: (a) they are two-fold differentially expressed genes (i.e. corresponding to mRNA with a change in expression of at least two-fold) at one time point of HGF stimulation with respect to control and (b) they were outputs of an interaction node of type transcription in the PID active-network.
    - Cell states: migration and proliferation, these nodes were chosen at the beginning of the study

Once this information is obtained, the following steps are required (Figure 1):

1. **Construction** of a computable graph from the PID NCI-Curated database, by using the XMLtoSIF [31] software.

2. **Filtering** of small molecules, such as: ATP, GTP, GDP, Calcium, IP3, DAG and subsequent removal of nodes related to *absent proteins*, i.e. those which were absent in both conditions measured in the microarray experiment.

3. **Pruning** of the network with respect to the seed nodes selection. This step consists of removal of nodes which are those that are not included in any path between any pair of seed nodes. Thus, the pruned graph is: $G' = (V', E') : (V', E') \in \sum_{s \neq t} \sum p_{s,t}$, where p_{st} is a path from node *s* to node *t*.

4. Run Monte Carlo random walk algorithm **MCWalk**, using as source and target nodes the designated seed nodes. Note that a node can belong to either or both sets. In this particular case all seed nodes are marked as both source and target nodes.

Monte Carlo random walks algorithm – MCWalk

The purpose of the MCWalk algorithm is: i) finding and quantifying the relationship of sets specified source and target nodes of interest (note that a node can belong in both sets), ii) ranking of the intermediate nodes that best explain this relationship and iii) extraction of relevant subgraphs based on this ranking. The frequency of traversal of the intermediate nodes expresses the probability that a node is traversed when a signal travels between source/target nodes. This is defined as random walk score *S*, identifying key regulator molecules that control signaling pathways in the network. Subsequently a family of subgraphs is obtained using a cutoff value *T* for this score. This approach allows us to obtain an adequate sample in the large space of all possible paths and networks between the nodes of interest.

The algorithm accepts as input two sets of nodes (source and target), and the graph *G(V,E)*, where the random walk is performed. At each run, a random walker starts from a source node and traverses the network in a random fashion. The random walk stops when it reaches a target node and the nodes and edges that have been traversed at least once are "marked" during this realization. The process is then repeated until the desirable number of runs is reached. The score *S* of each node at the end of the simulation is defined as the number of runs when this node is traversed at least once, over the total number of runs. This is proportional to the expected number of times this node is traversed when a signal is sent from any source to any target node (e.g. in a signal transduction pathway) and takes values $0 \leq s \leq 1$. The algorithm can be described in simple steps as follows:

1. The sets of source and target nodes are determined
2. A random source node is selected
3. At each time step the random walker moves into a random successor node and marks this node as traversed for the current run
4. If the successor node is a target node, then the process stops and one realization is complete.
5. The steps 2-4 are repeated until we reach the pre-assigned number of realizations, where the simulation stops.

After each node's score is assigned, a subgraph $G'(V', E') \subseteq G(V, E)$ can be extracted. This graph includes each node *u*, whose score S_u is above a specified

threshold *T*, such that $\forall u \in V', S_u \geq T$. Thus, *T* is used as a cutoff that controls which nodes will be included in a subgraph for a particular value of *T.* This score S_u is related to Newman's random walk betweeness [27], but applied in our case in directed graphs for specific paths of signal transduction, and describes each node's contribution to all paths connecting the source and target nodes. By relaxing *T* one can obtain a more complete picture in the pathways associating these nodes, at the expense of having a larger subgraph, while for large values only the most relevant nodes remain in the network resulting in a smaller, but more incomplete graph.

This approach can be combined with the shortest path methods in order to extract the underlying subgraph. This might be desirable, particularly for high threshold values, which tend to disconnect a large part of the network if we want to ensure that the seed nodes will remain connected. In this case, the nodes in the shortest paths between all source and target nodes are discovered with Dijkstra's algorithm [15] and are pre-assigned with a score of 1. Relaxing *T* adds more relevant nodes to the existing shortest path subgraph, combining the compactness of the shortest path method with the completeness of the random walk approach. Thus, the traversal of the network using the random walk between the seed nodes yields a family of subgraphs, with respect to the threshold cutoff value *T*.

The simulation is run on the pruned master network structure, which is critical to avoid paths that lead to "dead ends" (i.e. nodes with no successors, which are not target nodes), and therefore do not contribute in the exploration of the network. This pruning is also important to avoid a subtle bias effect; because these dead ends increase the probability to reach nodes which are near the source nodes compared to those which are further away, the signal will most likely get trapped in a dead end while trying to reach them. Note that this algorithm can be easily adapted to biased random walks on weighted networks, where the walker follows a successor with a probability equal to its edge weight (in our case we consider all weights equal to 1 and therefore the walk is completely random). If there are paths ending in non-seed nodes after the sub-networks are obtained, these can be removed in a final pruning step.

The random walk dependency matrix ***D*** quantifies the frequency with which a signal from a source node arrives at a target node. This measures the strength of relationship between these two nodes *u* and *v* based on the reachability of the target from the source i.e. the number of paths through which the source can influence the target. As a convention from hereon, the rows of the matrix represent the source and the columns the target nodes. $\boldsymbol{D}_{u,v}$ is then the relative frequency with which a signal from *u* reaches *v*, such as: $\forall u, \sum_v D_{u,v} = 1$.

This matrix can then be reordered based on the reachability of each node. First, all target nodes (columns) are reordered based on total number of times they are reached from all sources defined as: $\forall v, g(v) = \sum_u D_{u,v}$. Subsequently the source nodes are reordered based on how strongly they are connected with the most frequently accessed target nodes as: $\forall u,\ f(u) = \sum_v D_{u,v} g(v)$.

The MCWalk algorithm was implemented in Java, using the Java Universal Network/Graph library (JUNG - http://www.jung.sourceforge.net)

Results & Discussion

The PID active network is revealed to be scale-free, as shown by Figure 2. The degree distribution follows a power law of the form: $P(k) \propto k^{-\gamma}$, with γ=2.3 for the active network and γ=3 for the pruned network. γ is the degree exponent that controls (among others) the density of the network and the importance of highly connected nodes (hubs). Statistical large scale analysis of large scale metabolic [36] and protein interaction networks [37] revealed that such systems are described by scale-free networks. These deviate from the classical random network theory introduced by Erdös and Rènyi [38], in that their topology is extremely heterogeneous, dominated by a few highly connected nodes. This type of networks has been shown to be especially resilient under random failures [39], but also very prone to intentional attacks [40].

The pruned network is sparser than the active network and significantly smaller: 530 nodes instead of 9092 nodes in the original network. Therefore, while pruning reduces significantly the network complexity, it loosely retains the original scale free structure of the network (noise in the distribution tail is increased due to the small network size). This implies that our network is controlled by a small number of hubs, which are essential for its structure and will collapse if even a few of these nodes are removed. Examples of such nodes are ERK1/2 and uPAR. The network is very robust when nodes of low connectivity, such as Laminin6 are removed.

Correlation of random walk score and node degree

Typical centrality measures usually have a high correlation with the node degree. Betweeness centrality is known to be highly correlated with node degree in different types of networks [41]. Random walk centrality was also found to be highly correlated with degree in the Barabasi-Albert network [26]. Newman's random-walk

betweenness is moderately highly correlated with degree (R^2=0.626) and very highly correlated with shortest path betweenness (R^2= 0.923). Thus, in general, vertices with higher degree or higher shortest-path betweenness tend also to have higher random-walk betweenness [27].

In Figure 3 we examine the correlation of random walk score *S* (i.e frequency with which a node is traversed when a random signal travels between the seed nodes) and node degree *k*, where *k* is the total number of connections of each node. For a selection of random seed nodes, this correlation seems to be very weak (R^2= 0.37). This is even more the case when using case-specific seed nodes i.e. HGF receptor c-Met nodes, DE genes and cell states, and R^2 then drops to 0.07 (inset of Figure 2), indicating that the high degree nodes are not necessarily the ones controlling the signaling pathways in the network, as one would expect.

A significant number of nodes are found to have low *k* but high *S*, including effector molecules and precursors for biological processes. Such examples are the Laminin isoforms, which are important and biologically active parts of the basal lamina, influencing cell differentiation, migration and adhesion [42]. β-catenin, which is involved in Cysteine-rich angiogenic inducer 61 (Cyr61) regulation also falls in this category. Activation of β-catenin signaling elevates the mRNA level of Cyr61 in HepG2 cells, while inhibition of β-catenin signaling reduces both mRNA and protein levels of Cyr61 [43]. A smaller number of nodes, as expected from the scale free structure of the network, have high *S*, but also high *k,* indicating that they are key players in many biochemical processes. Such nodes are active RAC1/GTP, which stimulates endothelial cell migration and activates the JNK cascade reaction and alpha catenin, which participates in many biological complexes (see Table 1 for more such examples).

In order to ensure that all our seed nodes will remain connected, we set the score of all the shortest-path nodes discovered with Dijkstra's algorithm to 1. Thus, the minimal subgraph will contain the nodes only included on the shortest path and the maximal subgraph will be equivalent to the pruned network. The number of nodes/edges of this subgraph with respect to *T* is shown in Figure 4. Note that in the range in the range $0.001 < T < 0.1$ this number scales approximately as the logarithm of *T*. This is true both for the random and for the specified seed nodes selection set. As a result, we get a tradeoff between size and pathway completeness for the respective node set. The appropriate network size can then be selected according to the respective problem at hand. In our case we use a value of T=0.01 to highlight the differences between the random walk and shortest path algorithm (Figure 6).

Linking the seed nodes with a dependency matrix

In some cases it is useful to know even whether any influence between two nodes exists (i.e. a growth factor and a differentially regulated gene). For such questions standard methods from graph theory can be applied, such as the distance matrix, calculating the geodesic distance between two nodes (see e.g. [2] for such an application). While this matrix also allows us to see if two nodes are connected at all, provides little knowledge about the strength of the association. This also has little meaning in biological networks, because distance has very little meaning since information in such networks doesn't travel preferentially on shortest paths. As such, we attempt to quantify the relationship between the source and target nodes using the random walks by constructing the dependency matrix for all the seed nodes, shown in Figure 5.

The rows correspond to sources, i.e nodes where a randomly moving signal is produced the columns to the target nodes which receive this signal. The values of this matrix are the fraction of times that a signal travels from any source to any target node, while traversing the network randomly. Therefore, the random walk dependency matrix may provide a more realistic quantification of the relationship of two nodes. Note however that this is highly dependent upon the network structure and should only be considered as an indication rather than an absolute measure.

This approach has the advantage that it doesn't require any prior knowledge on gene expression data, as it relies only on the network structure. However, the disadvantage is that as it does not incorporate any functional information, the results can be somewhat biologically irrelevant. To address this issue, an expression dependency matrix can be employed, which relies in the microarray expression data. This has the advantage that the reordering of the targets (e.g. differentially expressed genes) is done according to: $g(v) = \left|\log_2 f_c(v)\right|$, where f_c is the fold change of *v* obtained in the microarray. The source nodes are then reordered according to the above formula, which expresses how correlated they are (in terms of frequency of reachability) with the most strongly expressed differentially regulated genes. In this sense, the expression dependency matrix allows a more biologically relevant picture in the context of growth factor signal transduction pathways.

By examining the dependency matrix, it is possible to obtain information about the strength of relationship between our seed nodes. Laminin-alpha 3, which is exclusively linked with cell migration, is known to induce keratinocyte migration, playing an important role in re-epithelialization at tissue remodeling [44]. uPAR is strongly connected with cell migration; uPAR-knockdown cells also display greatly reduced migration and invasion rates, as well as a complete loss of the cells' ability to augment their invasiveness following plasminogen supplementation [45]. Met is

also found to be significantly associated with proliferation and migration. Met controls cell migration and growth in embryogenesis; it also controls growth, invasion and metastasis in cancer cells; and activating Met mutations predispose to human cancer [46]. Met is known to be essential for wound healing, as its signaling not only controls cell growth and migration during embryogenesis but is also essential for the generation of the hyper-proliferative epithelium in skin wounds, and thus for a fundamental regenerative process [47].

Pathway analysis of the MCWalk extracted network

In Figure 6 we present a version of the HGF stimulation network obtained using the random walk algorithm. Nodes which are included in the shortest path nave a black border, while nodes that are only obtained with the random walk traversal are marked in red to highlight the differences in subgraph extraction between of the two algorithms. Note that two or more nodes may share the same name. This is so because short names are used as labels for the nodes. These correspond to unique identifiers which include information about different modifications (e.g. phosphorylation information) for every node [31], but short names are shown for simplicity in visualizing the network. Many uPAR complex members, such as UPAR/PAI-1/Integrin and pro-uPA/uPAR are only available in the random walk version of the network.

The Neural Wiskott Aldrich Syndrome Protein (N-WASP) pathway which activates cell migration is also absent in the shortest path version. N-WASP-/- cells were found to migrate more rapidly than N-WASP+/+ cells in a scratch migration assay, suggesting that N-WASP deficiency leads to reduced adhesion to fibronectin and increased cell motility [48]. The NF-kappaB pathway is another important omission of the shortest path method. The transcription factor NF-kappaB is activated in response to a wide variety of stimuli, including growth factors, and is involved in biological responses in part overlapping with those triggered by HGF [30]. NF-kappaB is also involved in the Jun-N-terminal kinase and NF-kappaB pathways in the repression of the human COL1A2 gene [49].

Further examples involve pathways such as the uPA/Plasmin system mediated MMP-9 activation. Here, shortest path just discovers the direct connection from uPA to endothelial cell migration. It has been shown however, that uPA is involved in human bronchial epithelial cells migration, and this action is mediated by the generation of plasmin, which in turn activates MMP-9, thus making possible cell migration [50]. Moreover, we get a more complete picture of the uPA activation from RACO-1, which is a co-activator that links c-Jun to growth factor signaling and is essential for AP‑1 function in proliferation [51]. Other signaling pathways involve

cadherin clustering regulation, which is a key determinant of strong cell–cell adhesion [52]. It has been demonstrated that p120 catenin is required for growth factor–dependent cell motility and scattering in epithelial cells [53]. It was further suggested that in fibroblasts p120 could affect other cell functions such as organization of the actin cytoskeleton and cell motility by virtue of its influence on the activity of Rho GTPases [54–56].

Conclusions

We have developed a method (MCWalk) based on random walks using Monte Carlo computer simulations to connect input to output nodes, associate target genes and cellular processes and extract relevant subgraphs that best capture their relationship. This method can be used in any network (directed/undirected) and with no restrictions on the selection of the source and target nodes. It ranks all the intermediate nodes based on their random walk traversal, highlighting key regulator nodes in the network, and allows the extraction of subgraphs with respect to that score. We have applied this method on an example dataset of HGF stimulation in a keratinocyte-fibroblast co-culture [29] to retrieve a global network from NCI Pathway Interaction Database.

We have shown that the random walk score in the studied sample network is very weakly correlated with the node degree, contrary to what might be expected. In addition, we obtained a dependency matrix based on the frequency of random walk traversal between the seed nodes, as more reasonable alternative to standard methods, such as the distance matrix. We have also presented an expanded version of the network and highlighted important differences with standard shortest path methods. This method complements our previously published work [31] consisting of fully automated steps to construct case-specific signaling networks, combining microarray data and online databases. Therefore, it may provide a more complete approach in analyzing large-scale biological networks and retrieval of case-specific relevant subgraphs, linking sets of input and output nodes into coherent signaling networks with a distinct underlying biological function.

References

[1] Q. A. Soltow, D. P. Jones, and D. E. L. Promislow, Integrative and Comparative Biology **50**, 844 (2010).

[2] S. Klamt, J. Saez-Rodriguez, J. A. Lindquist, L. Simeoni, and E. D. Gilles, BMC Bioinformatics **7**, 56 (2006).

[3] L. Lovász, Combinatorics **2**, 1 (1993).

[4] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences of the United States of America **105**, 1118 (2008).

[5] K. Komurov, M. A. White, and P. T. Ram, PLoS Computational Biology **6**, (2010).

[6] G. H. Weiss, *Aspects and Applications of the Random Walk* (North-Holland, Amsterdam, 1994).

[7] A. Kittas, S. Carmi, S. Havlin, and P. Argyrakis, EPL (Europhysics Letters) **84**, 40008 (2008).

[8] A. Kittas and P. Argyrakis, Physical Review E **80**, (2009).

[9] L. Gallos, Physical Review E **70**, 1 (2004).

[10] L. Gallos and P. Argyrakis, Physical Review Letters **92**, 1 (2004).

[11] S. Weber, M.-T. Hütt, and M. Porto, EPL (Europhysics Letters) **82**, 28003 (2008).

[12] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics **71**, 056104 (2005).

[13] K. Faust, P. Dupont, J. Callut, and J. van Helden, Bioinformatics (Oxford, England) **26**, 1211 (2010).

[14] J. van Helden, A. Naim, R. Mancuso, M. Eldridge, L. Wernisch, D. Gilbert, and S. J. Wodak, Biological Chemistry **381**, 921 (n.d.).

[15] E. W. Dijkstra, Numerische Mathematik **1**, 269 (1959).

[16] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, Nucleic Acids Research **30**, 1575 (2002).

[17] W. S. Verwoerd, BMC Systems Biology **5**, 25 (2011).

[18] R. Guimerà and L. A. Nunes Amaral, Nature **433**, 895 (2005).

[19] P. Dupont, J. Callut, G. Dooms, J.-N. Monette, and Y. Deville, Research Report UCL/FSA/INGI RR 2006-07 (n.d.).

[20] S. P. Borgatti and M. G. Everett, Social Networks **28**, 466 (2006).

[21] L. C. Freeman, Social Networks **1**, 215 (1979).

[22] B. H. Junker and F. Schreiber, editors, *Analysis of Biological Networks* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008).

[23] B. H. Junker, D. Koschützki, and F. Schreiber, BMC Bioinformatics **7**, 219 (2006).

[24] L. C. Freeman, Sociometry **40**, 35 (1977).

[25] P. Bonacich, The Journal of Mathematical Sociology **2**, 113 (1972).

[26] J. D. Noh, Physical Review Letters **92**, (2004).

[27] M. E. J. Newman, Social Networks **27**, 39 (2005).

[28] E. Estrada, D. J. Higham, and N. Hatano, Physica A: Statistical Mechanics and Its Applications **388**, 764 (2009).

[29] H. Busch, D. Camacho-Trullio, Z. Rogon, K. Breuhahn, P. Angel, R. Eils, and A. Szabowski, Molecular Systems Biology **4**, 199 (2008).

[30] M. Müller, A. Morotti, and C. Ponzetto, Molecular and Cellular Biology **22**, 1060 (2002).

[31] C. Guziolowski, A. Kittas, F. Dittmann, and N. Grabe, The FEBS Journal **279**, 3462 (2012).

[32] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, Nucleic Acids Research **37**, D674 (2009).

[33] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, Bioinformatics (Oxford, England) **27**, 431 (2011).

[34] D. E. Knuth, Science (New York, N.Y.) **194**, 1235 (1976).

[35] Y.-K. Shih and S. Parthasarathy, Bioinformatics (Oxford, England) **28**, i49 (2012).

[36] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, Nature **407**, 651 (2000).

[37] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, Nature **411**, 41 (2001).

[38] B. Bollobas, Transactions of the American Mathematical Society **286**, 257 (1984).

[39] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, Physical Review Letters **85**, 4626 (2000).

[40] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, Physical Review Letters **86**, 3682 (2001).

[41] K.-I. Goh, E. Oh, B. Kahng, and D. Kim, Physical Review E **67**, 017101 (2003).

[42] R. Timpl, H. Rohde, P. G. Robey, S. I. Rennard, J. M. Foidart, and G. R. Martin, The Journal of Biological Chemistry **254**, 9933 (1979).

[43] Z.-Q. Li, W. Ding, S.-J. Sun, J. Li, J. Pan, C. Zhao, W.-R. Wu, and W.-K. Si, PloS One **7**, e35754 (2012).

[44] Y. Momota, N. Suzuki, Y. Kasuya, T. Kobayashi, M. Mizoguchi, F. Yokoyama, M. Nomizu, H. Shinkai, T. Iwasaki, and A. Utani, Journal of Receptor and Signal Transduction Research **25**, 1 (2005).

[45] T. S. Nowicki, H. Zhao, Z. Darzynkiewicz, A. Moscatello, E. Shin, S. Schantz, R. K. Tiwari, and J. Geliebter, Cell Cycle (Georgetown, Tex.) **10**, 100 (2011).

[46] C. Birchmeier, W. Birchmeier, E. Gherardi, and G. F. Vande Woude, Nature Reviews. Molecular Cell Biology **4**, 915 (2003).

[47] J. Chmielowiec, M. Borowiak, M. Morkel, T. Stradal, B. Munz, S. Werner, J. Wehland, C. Birchmeier, and W. Birchmeier, The Journal of Cell Biology **177**, 151 (2007).

[48] A. Misra, R. P. Z. Lim, Z. Wu, and T. Thanabalu, Biochemical and Biophysical Research Communications **364**, 908 (2007).

[49] F. Verrecchia, E. F. Wagner, and A. Mauviel, EMBO Reports **3**, 1069 (2002).

[50] C. Legrand, M. Polette, J. M. Tournier, S. de Bentzmann, E. Huet, M. Monteau, and P. Birembaut, Experimental Cell Research **264**, 326 (2001).

[51] C. C. Davies, A. Chakraborty, F. Cipriani, K. Haigh, J. J. Haigh, and A. Behrens, Nature Cell Biology **12**, 963 (2010).

[52] P. Z. Anastasiadis and A. B. Reynolds, Journal of Cell Science **113 (Pt 8**, 1319 (2000).

[53] M. Cozzolino, V. Stagni, L. Spinardi, N. Campioni, C. Fiorentini, E. Salvati, S. Alemà, and A. M. Salvatore, Molecular Biology of the Cell **14**, 1964 (2003).

[54] N. K. Noren, B. P. Liu, K. Burridge, and B. Kreft, The Journal of Cell Biology **150**, 567 (2000).

[55] P. Z. Anastasiadis and A. B. Reynolds, Current Opinion in Cell Biology **13**, 604 (2001).

[56] I. Grosheva, M. Shtutman, M. Elbaum, and A. D. Bershadsky, Journal of Cell Science **114**, 695 (2001).

Figures

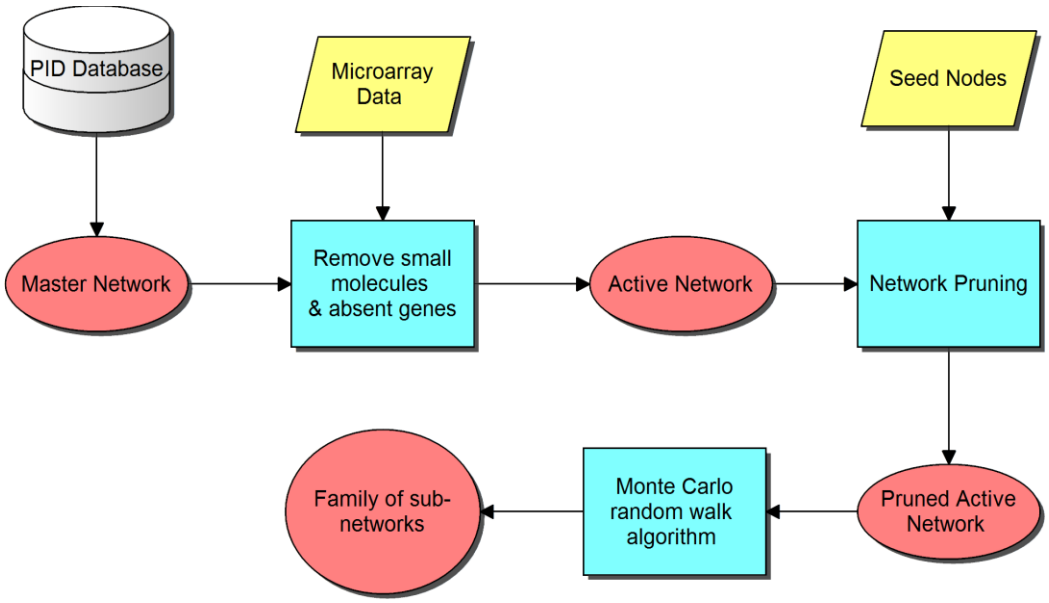

Figure 1 - Pipeline describing the steps required to integrate automatically the regulatory knowledge stored in the Pathway Interaction Database and the biological observations given a microarray experiment. The networks obtained can be read in Cytoscape. Blue boxes correspond to automatic steps for data filtering and preprocessing. The pruned active network is used an input for the Monte Carlo random walk algorithm which obtains a family of subnetworks depending on the score threshold *T*.

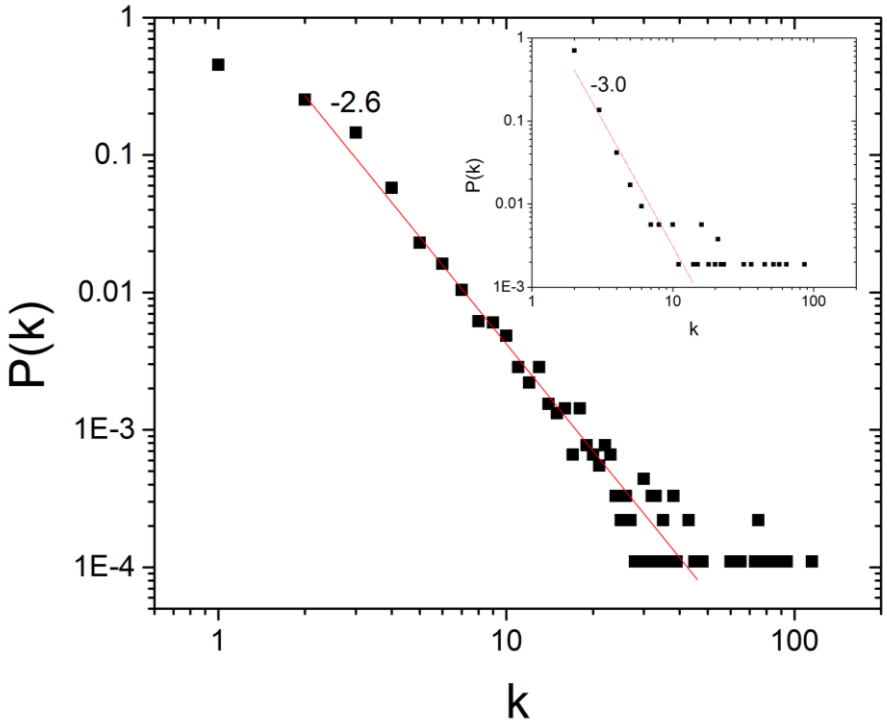

Figure 2 - Normalized degree distribution *P(k)* of the PID active network (*x* and *y* axes are logarithmic.). Inset: Same, but for the pruned active network. *k* is the total degree of the node (number of links) and *P(k)* the probability distribution of *k* in the network. Red lines correspond to a power law $P(k) \propto k^{-\gamma}$, with γ=2.6 for the active network and γ=3 for the pruned active network. Thus, the network has a scale-free structure dominated with a few highly connected nodes (hubs) and a large number of nodes with very few links.

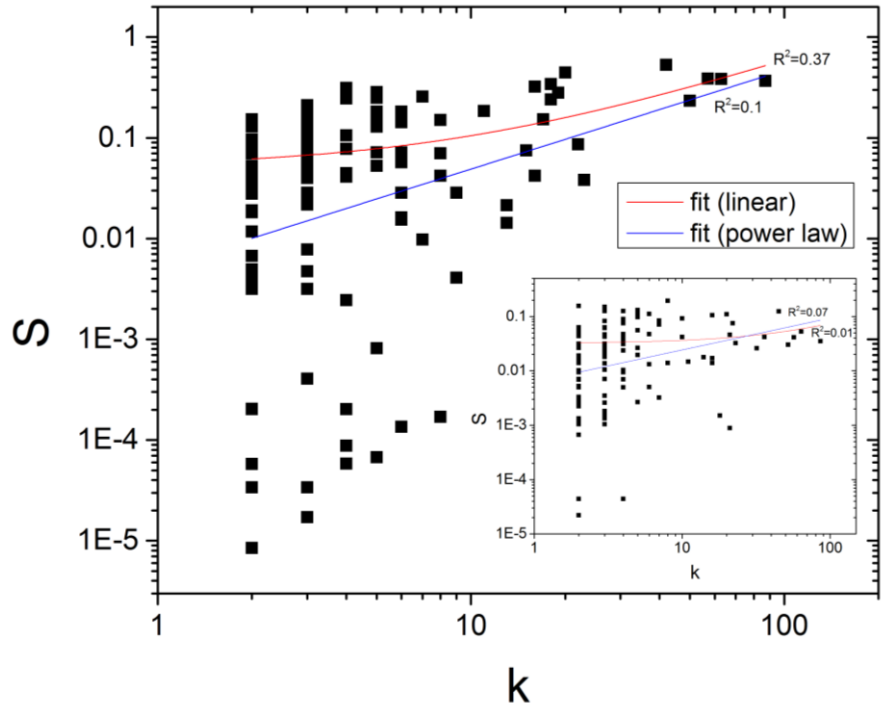

Figure 3 – Random walk score *S* vs degree *k* for 50 random seed nodes. Inset: Same but using the specified seed nodes (i.e. HGF receptor c-Met nodes, DE genes and cell states). *x* and *y* axes are logarithmic. Lines correspond to linear and power law fitting. Correlation between score and degree is very weak, especially for our specific case of seed nodes selection (inset), in contrast to popular measures such as betweeness or closeness centrality.

Name	Degree (*k*)	Score (*S*)
uPA	8	0.19515
alpha6/beta4 Integrin/Laminin[l]	2	0.15539
HGF/MET	3	0.15017
alpha6/beta4 Integrin/Laminin[h]	5	0.13081
HGF(dimer)/MET(dimer){+}[itm]	4	0.125
Laminin	3	0.12499
RAC1/GTP{+}	45	0.12386
cell adhesion	6	0.11178
E-cadherin/beta catenin/alpha catenin[bpm]	5	0.10971
alpha catenin[c]	20	0.10932
E-cadherin(dimer)/Ca2+[cj]	16	0.10505
E-cadherin/beta catenin/alpha catenin[cj]	5	0.09711
AP1{+}[n]	10	0.09227
E-cadherin/Ca2+/beta catenin/alpha catenin[cj]	4	0.08893
JUN/FOS{+}[n]	4	0.08513
E-cadherin/Ca2+/gamma catenin/alpha catenin/p120 catenin[cj]	4	0.08457
uPA/uPAR (dimer){+}[pm]	7	0.08334
E-cadherin/gamma catenin/alpha catenin[bpm]	3	0.08232
ARF6/GTP{+}[pm]	22	0.07508

Table 1 – List of top 20 ranked intermediate nodes in the HGF network based on their random walk score *S* (see Figure 6 for corresponding legend in name symbols).

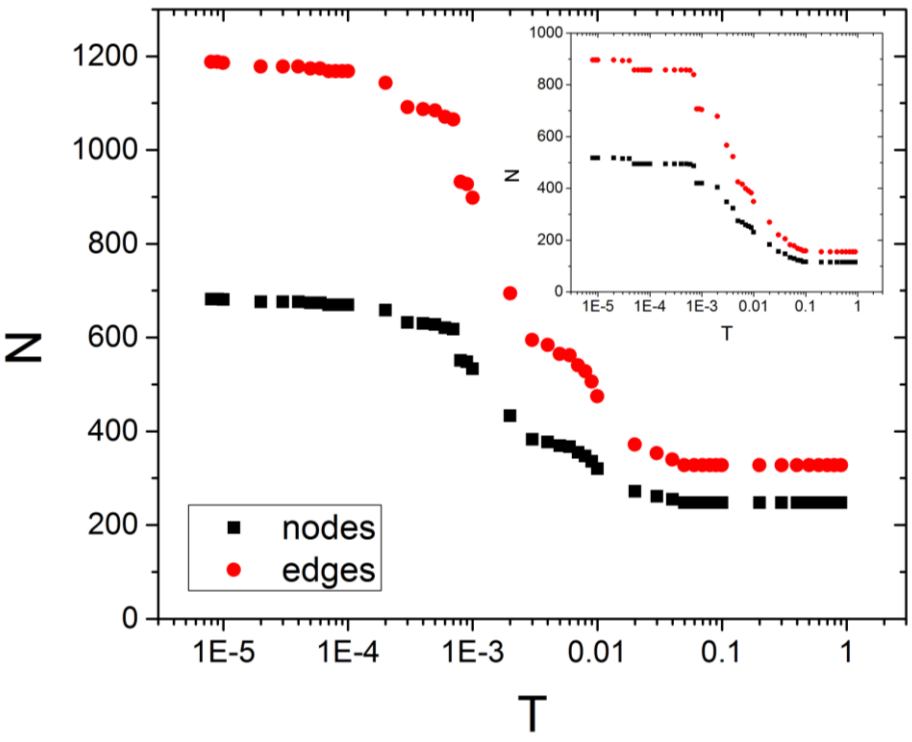

Figure 4 – Number of nodes and edges vs random walk threshold *T* for 50 random seed nodes. Inset: Same but using the designated seed nodes (i.e. HGF receptor c-Met nodes, DE genes and cell states). *x* axis is logarithmic. High threshold values include only nodes that are in the shortest path (these nodes are preassigned with a score of 1 and thus always present). The relation is close to logarithmic in the range $0.001 < T < 0.1$.

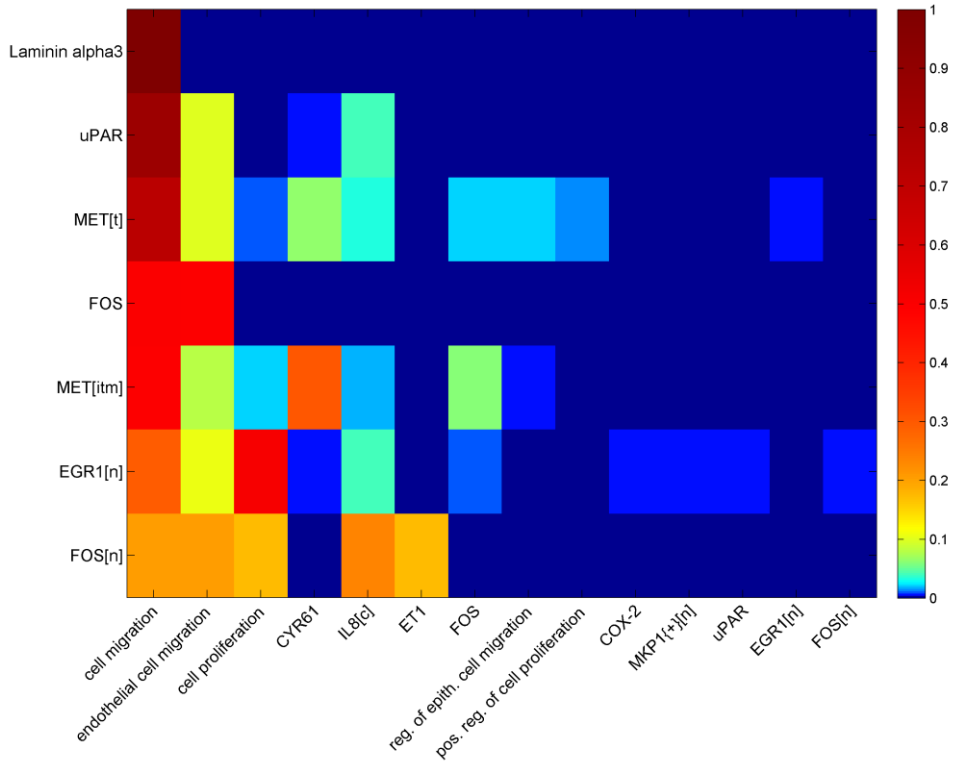

Figure 5 - Dependency matrix linking the chosen seed nodes. Column (targets) reordering is done based on the reachability of each node. Values in each matrix element are corresponding probabilities for a signal to arrive from a specific source to a target node. The rows (sources) are then reordered based on how strongly they are connected with the most frequently accessed target nodes. Color scale is logarithmic.

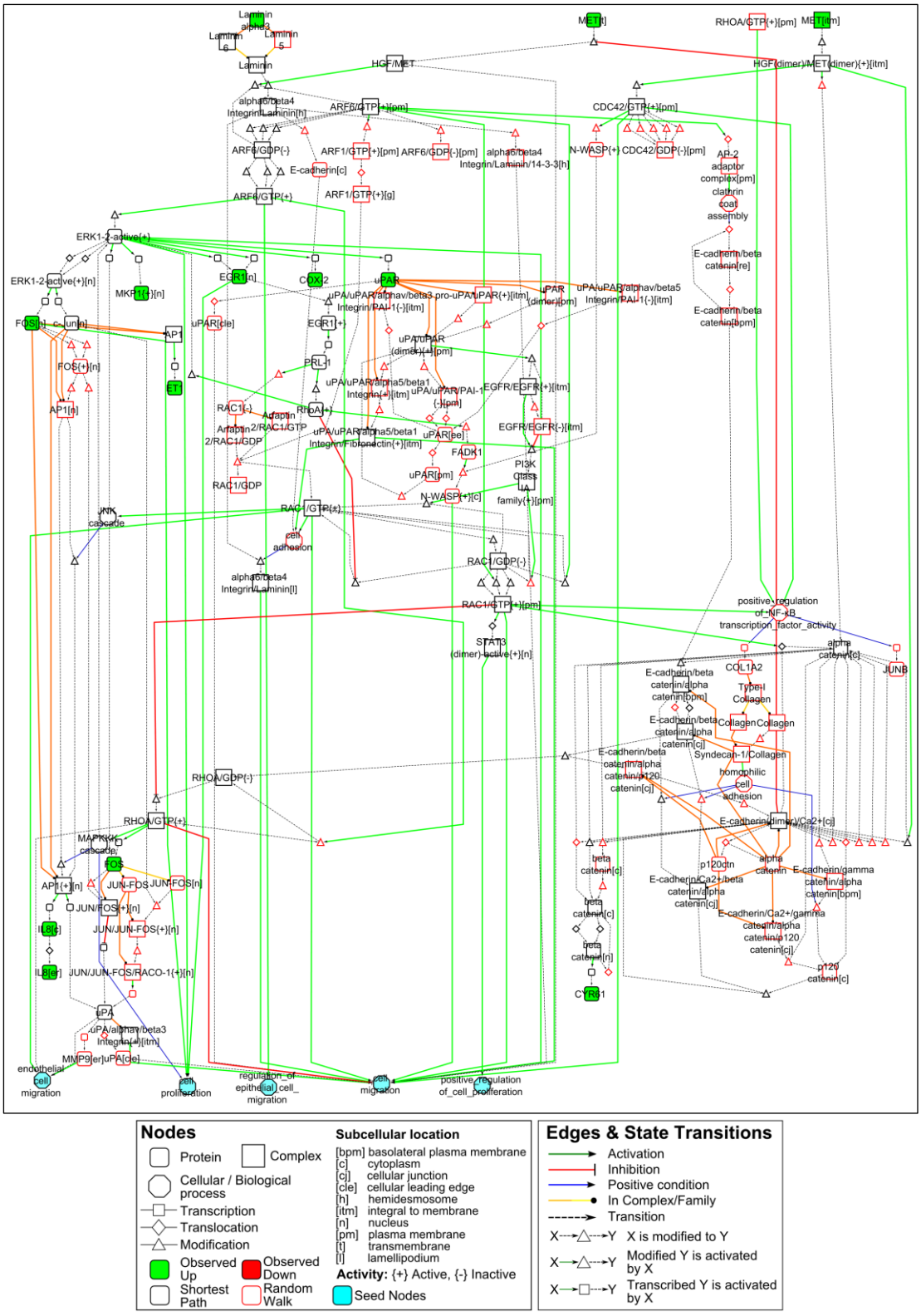

Figure 6 - Network obtained using the MCWalk algorithm for threshold *T* = 0.01 using as seed nodes the HGF receptor c-Met nodes, DE genes and cell states after 1hr of HGF stimulation. Nodes with black border are obtained using Dijkstra's shortest path algorithms, while nodes with red border are obtainable only with the Monte Carlo random walk algorithm. Number of nodes/edges with i) shortest path: 115/155, ii) random walk (*T* = 0.01): 231/349. Short names of nodes are shown on the network; two nodes with the same name may have different modifications (e.g. phosphorylation) and are represented in the network with a unique identifier.