단백질 상호작용 네트워크의 삼각형 기반 변 점수 산정법

2017년

전 현 성

교육학석사학위논문

단백질 상호작용 네트워크의 삼각형 기반 변 점수 산정법

A triangle based edge scoring of protein interaction network

2017년 8월

서울대학교 대학원

수학교육과

전 현 성

단백질 상호작용 네트워크의 삼각형 기반 변 점수 산정법

지도 교수 김 서 령

이 논문을 교육학석사 학위논문으로 제출함
2017년 7월

서울대학교 대학원
수학교육과 수학교육전공
전 현 성

전현성의 교육학석사 학위논문을 인준함
2017년 8월

위 원 장 _____ 조 한 혁 _____ (인)

부위원장 _____ 유 연 주 _____ (인)

위 원 _____ 김 서 령 _____ (인)

Abstract

Motivation: Uncovering the mystery of evolutionary mechanism of protein interaction networks has been actively conducted in order to understand interactions of proteins that induce biological processes in organisms. There have been many attempts to solve the mystery by proposing evolutionary models of protein interaction networks. Topological properties of protein interaction networks are mentioned several times and given an important role in these attempts since a validation of suggested models is made through topological properties of known protein interaction networks. While one group of researchers have made efforts to generate current protein interaction networks from some hypothetical infant state of protein interaction networks through suggested evolutionary models, another group of researchers have made efforts to estimate the phylogenetic age of proteins from evolutionary relationships. Recently, these efforts gave rise to the database of phylogenetic age of proteins and this allows many researchers to estimate ages of proteins in their interest easily. Recent studies on Mendelian diseases and cancer suggested that proteins associated with specific diseases populate certain category of the phylogenetic age of proteins.

The fact that the topological properties of the protein interaction network have played important roles in the evolution of protein interaction networks tells us that topological properties of protein interaction network and properties of proteins, which is related to the evolution of the protein interaction network, is closely related in some level.

As one can see from closeness in terms, the evolutionary model of protein interaction networks and phylogenetic age of proteins are closely related and thus topological properties of protein interactions, which is important in studies of the evolutionary models, can be used to estimate the phylogenetic age of proteins. Besides, the research results on the relationship between diseases and phylogenetic age of proteins motivate us to predict proteins associated to diseases by utilizing topological properties of protein interaction networks.

Results: We construct a weighted human protein interaction network from a human protein interaction network which is provided via BioGRID database. The weight of an edge is defined as the number of triangles which contains this edge in the protein interaction network and thus we call this weight as the triangle score. We make comparison between the edge scores of a human protein interaction network given by STRING database and the triangle score. In an

attempt to find relationship between the triangle score and properties of proteins that is related to the evolution of protein interaction networks, we make comparison between the triangle score and bit score, which is a measurement of protein sequence similarity. Moreover, we attempt to sieve out self-interacting proteins from the whole human proteins based on the triangle score. In an effort to predict the phylogenetic age of proteins based on the triangle score, firstly, we extract proteins that are incident on an edge that has a high triangle score from the weighted protein interaction network which we constructed with the triangle score. After the extraction, we make inquiries to the *ProteinHistorian* database to get phylogenetic ages of extracted proteins. Finally, we show that there is a relationship between triangle score and phylogenetic age by comparing the ratio of proteins with each phylogenetic age to whole human proteins and the ratio of extracted proteins with each phylogenetic age to whole extracted proteins. Based on the triangle score, we also attempt to predict disease associated proteins for several diseases.

Keywords: protein interaction network; weighted protein interaction network; disease associated protein; phylogenetic age; heterogeneous network

Student Number: 2015-21610

목 차

그림 목차

제 1 장

Introduction

Proteins play diverse and crucial roles in biological tissues as main catalysts, signaling messengers and molecular machines. These biological processes, through which meaningful life forms exist, are regulated by physical contacts among proteins and these physical contacts are called protein interactions. Thus protein interactions are vital to understanding protein functions and biological processes. Recently, increasing number of protein interactions have been detected due to high-throughput data and advances in experimental methods. Moreover, bioinformatics methods or computational methods, which have been developed to overcome time-consuming and costly nature of experimental methods, have paved a way to faster and easier predictions of protein interactions.

While high-throughput data and advances in experimental methods aid researchers in reporting enough protein interactions to construct several number of protein interaction databases, the reliability of the reported protein interactions is doubtful. To banish the doubt on the reliability of reported proteins, researchers have

tried to measure the reliability of given protein interactions. One of the earliest attempt at this direction is made by Nabieva et al. [Nabieva et al. (2005)]. In [Nabieva et al. (2005)], edge scoring method based on the experimental sources that contribute to the given protein interactions is suggested. To determine the score, firstly, they separated all the experimental sources that contribute to the physical interaction data into several groups. After this grouping, they computed the fraction of interactions which connect proteins with a known shared function for each group. Finally, they defined reliability of the edge by $1-\prod_i(1-r_i)$, where r_i is the computed fraction of group i and the product is taken over all experiments i where this interaction is found. In [Szklarczyk et al. (2010)], integrated form of repository with scored edges, called STRING, is compiled. The edge score is derived by separately benchmarking groups of associations against the manually curated functional classification scheme of the KEGG database. Each score represents a rough estimate of how likely a given association describes a functional linkage between two proteins that is at least as specific as that between an average pair of proteins annotated on the same 'map' or 'pathway' in KEGG.

With a vast number of reported or predicted protein interactions at their hands, researchers begin to see protein

interactions at the network level and this new perspective about protein interactions leads to protein interaction networks databases such as BioGRID, the Human Protein Reference Database and Database of Interacting Proteins. Studies on protein interaction networks revealed that protein interaction networks possess properties that are not present in randomly generated networks. One example of the property is the degree distribution. The degree distribution is the fraction of nodes with a given degree and the degree distribution of a random network follows Poisson distribution. In contrast, the degree distribution of a protein interaction network follows power law tail distribution, thus the fraction of nodes with degree d is approximately $Cd^{-\gamma}$ for some constant C and γ. The networks which have the degree distribution of power law tail are called scale−free networks and most noticeable consequence of having the degree distribution of power law tail is the presence of a few highly connected nodes that holds the network together, this highly connected nodes are called hubs. These peculiar properties possessed by protein interaction networks lead researchers to ask the natural question: How the current protein interaction networks have been formed through evolution?

Regarding the question, several mechanisms, by which protein interaction networks evolve, have been suggested. Most researchers believe that gene duplication is one of the mechanisms and post-duplication divergence also plays some role in the evolution of protein interaction networks. Gene duplication is regarded as the primary evolutionary phenomenon which drives protein network growth and it is represented in protein interaction networks as a node and edge duplication. Post-duplication divergence refers to the events that protein interaction networks go through after gene duplication and researchers believe that one of a prevalent event that occurs after gene duplication is subfunctionalization, that is a deletion of redundant interactions from gene duplication. After gene duplication, protein interaction networks possess redundant interactions and one of these interactions can be deleted from the networks without obstructing crucial cellular functions since redundant interactions play exactly the same role in a cellular organism. While most researchers agree that gene duplication and post-duplication divergence are the methods by which protein interaction networks evolve, they do not have a consensus on whether the method called neofunctionalization is one of the main driver of the evolution of protein interaction networks. One of the earliest evolutionary model of protein

interaction networks has a high rate of neofunctionalization to explain an irregularly high number of triangles in protein interaction networks than a number of triangles in randomly generated scale free networks [Sole et al. (2002)]. However, Vazquez et al. suggested an evolutionary model in which neofunctionalization never happens [Vazquez et al. (2003)]. In this model new interactions are formed through duplication of self-interacting proteins. Later Gibson et al. claimed that neofunctionalization is erroneously identified as a significant factor in evolutionary models of protein interaction networks. To prove their claim, Gibson et al. showed that the evolutionary model suggested by Vazquez et al., which is the evolutionary model without neofunctionalization, can generate enough triangles in the simulation of generating protein interaction networks under the assumption that self-interacting proteins are underreported in current studies. Moreover, they suggested an evolutionary model that considers interaction sites on protein surface as the subfunctionalization unit rather than each interaction [Gibson et al. (2011)]. Though Gibson et al. showed that current protein interaction networks can be generated without neofunctionalization, researchers still cannot agree upon whether neofunctionalization occurs during the evolution and Peterson et al. suggested an evolutionary model with neofunctionalization that

considers interaction surface on protein [Peterson et al. (2012)]. While the question on which methods should be considered as main drivers in the evolution of protein interaction networks still cannot be answered with certainty, studies on evolutionary models showed us that triangles in protein interaction networks play important role in validating evolutionary models.

With this importance of the triangles in protein interaction networks in their minds, Gibson et al. assessed the validities of the evolutionary models of protein interaction networks based on the clustering coefficient, the value of which depends on the number of triangles in a protein interaction network, in their work. The *clustering coefficient* C for a given graph is defined as $\mathrm{C} = 3T/\Gamma$, where T is the number of triangles in the graph and Γ is the number of connected triples in the graph. The clustering coefficient is a relevant measure of the validity of an evolutionary model since gene duplication, subfunctionalization, and neofunctionalization each produces a measurable change in the number of triangles and the number of connected triples which comprise the clustering coefficient [Gibson et al. (2009)].

Another kind of studies considering the evolution of proteins has concentrated on an estimation of the phylogenetic age of proteins. Phylogenetic age of proteins represents the evolutionary

history of proteins and it is closely related to the functional history of proteins. Recently, researchers have found a relationship between certain evolutionary signatures and Mendelian diseases. Researchers also reported that cancer is related to the emergence of multicellularity during the evolution [Domazet-Loso et al. (2010)]. Since diseases are closely related to the functions of proteins, a lot of research relating diseases with evolutionary trait clearly shows that phylogenetic age of proteins is related to protein functions. Motivated by the relationship between phylogenetic age and protein functions, Capra et al. developed a software that estimates a given list of proteins based on a species tree, a protein family database and an ancestral family reconstruction [Capra et al. (2012)].

As most studies in biology deal with human health issues, there has been a lot of research trying to utilize protein interaction networks in drug studies and disease studies. One of the try outs in this direction is an effort to understand diseases at the network level rather than treat diseases as a consequence of an abnormality in a single gene. The knowledge composed by this kind of attempts is referred as network medicine and it is comprehensively described in the work of Albert-Laszio Barabasi et al. [Albert-Laszio Barabasi et al. (2011)]. The research on network medicine

is heavily rely on one particular property of biological networks and that property is the presence of modules in biological networks. The modules are highly interlinked local regions in the network and the term ‘highly linked’ is defined differently for each studies depending on contexts. Other than the properties of biological networks, network medicine is also based on hypotheses that connects structure of network to biological function and disease. The hypotheses of network medicine are as follows [Albert-Laszio Barabasi et al. (2011)]:

The first hypothesis is for the hubs in the network, network medicine assumes that non-essential disease genes segregate at the functional periphery of the interactome. In the embryo state, essential genes are associated with hubs.

The second hypothesis is called local hypothesis and it states that proteins involved in the same disease are more likely to interact with each other. If this hypothesis is true, then mutation in interacting proteins often lead to similar disease phenotype.

The third one is called disease module hypothesis. It says that cellular components associated with a specific disease phenotype are clustered in the same network neighborhood.

The fourth one is related to biological pathways and called network parsimony principle. Network medicine assumes that

causal molecular pathways of a disease often coincide with the shortest molecular paths between known disease associated components.

The last one is called shared component hypothesis and it tells us that diseases that share disease associated components are phenotypically similar.

Overall, researches of network medicine claims that disease associated components have noticeable topological properties of biological networks and make efforts to prove the claim.

Researches that utilize protein interaction networks to find gene-disease relationships naturally led researchers to find gene-phenotype relationships since elucidating the inherited basis of human disease involves linking genomic variation to clinical phenotype [Li et al. (2010)]. Recently, Li et al. suggested a gene-phenotype inferring algorithm which outperforms past algorithms by connecting a protein interaction network with a phenotype network and simulating random walk with restart on this networks, they call this network as a heterogeneous network [Li et al. (2010)]. This research shows that connecting protein interaction networks with other biological networks can yield a better result than working with protein interaction networks alone.

In this paper, we construct weighted protein interaction networks based on triangles in protein interaction networks and make comparison between this weights and weights provided by existing database. We attempt to find a relationship between the weights based on the triangles and properties of proteins related to the evolution of protein interaction networks. Furthermore, we show that topological properties, especially triangles in protein interaction networks, are related to the phylogenetic age of proteins. Motivated by network medicine researches, we attempt to predict disease associated proteins for several diseases based on triangles in protein interaction networks. Moreover, we construct a heterogeneous network which connects a weighted protein interaction network, weighted via triangle scores, and a phenotype network and then perform a random walk with restart on it to infer gene−phenotype relationships. After this, we compare our result with the result from the previous works.

제 2 장

Materials and Methods

In this chapter, we explain some biological databases that are used in this paper and introduce biological terms and graph theoretical terms that are necessary to understand our methods and results.

BioGRID database

BioGRID is one of protein interactions databases that provides public access to its data. The BioGRID was firstly released as the GRID (the general repository for interaction datasets) in 2003 [Breitkreutz et al. (2003)]. However, this name causes some confusion since there was several GRID computing projects with almost identical name. So, they changed the name of the database to the BioGRID. The BioGRID was initially developed by the Lunenfeld-Tanenbaum Research Institute at Mount Sinai Hospital but now the BioGRID team includes many institutes around The united states and Canada such as the institut de Recherche en Immunologie et en Cancerologie at the Universite de Montreal, the Lewis-Sigler Institute for Integrative Genomics at Princeton University, and the Wellcome Trust Center for Cell Biology at the

University of Edinburgh. At the first stage of its development, the BioGRID mainly focused on curation of binary protein−protein and genetic interactions, but through many updates [Stark et al. (2006); Breitkreutz et al. (2008); Stark et al. (2011); Chatr−Aryamontri et al. (2013); Chatr−Aryamontri et al. (2015); Chatr−Aryamontri et al. (2017)], it now holds curated post−translational modification data, chemical reaction data, and complex multi−gene/protein interactions. Currently, the BioGRID archives protein interaction data from model organisms and humans. The BioGRID holds over 1,400,000 interactions curated from high−throughput datasets and individual studies. It focuses on curating protein interactions that is conserved through the evolution and relevant to human health and it is updated on a monthly basis. We use protein interaction data of Homo sapiens. Though current version of BioGrid database is 3.4.148, in this paper, we use the database of version 3.4.144.

Protein interaction network

As mentioned in the introduction, protein interactions are important to understand how cell components work and how actual cellular processes happen. At the cellular level, protein interactions are represented by the protein interaction network, which is the union of all proteins and the interactions among them since many protein interactions contribute some functionality to one or more

cellular structures or cellular processes [Vazquez (2010)]. An example of the protein interaction network is shown in figure 2-1.

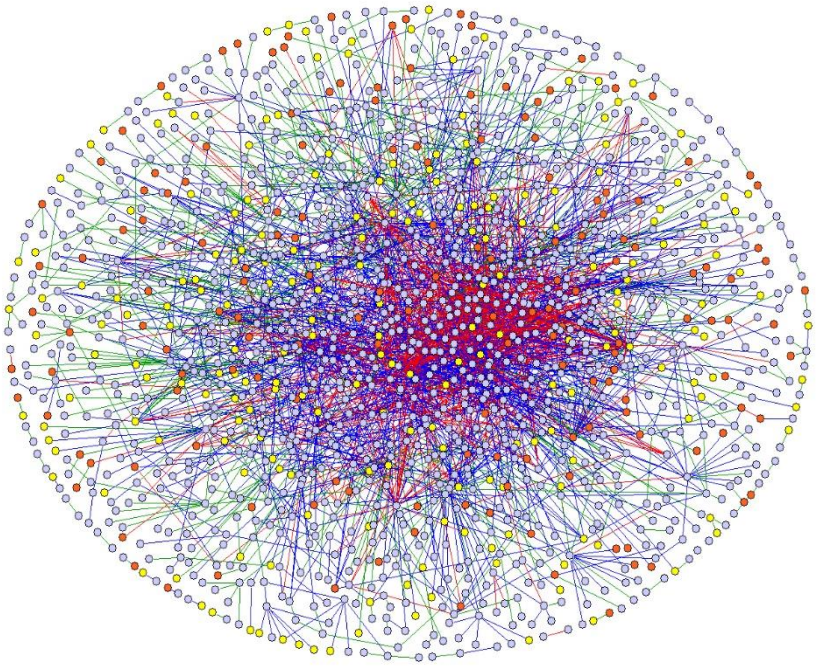

Figure 2-1. Human protein interaction network

Figure 2-1 is one of the first map showing human protein interactions and it is mapped by the scientists of Max Delbruck Center for Molecular Medicine [Stelzl et al. (2005)]. This human protein interaction network shows 3,200 protein interactions between 1,700 proteins. However, nowadays, researchers estimate the whole human protein interaction network would have 650,000 interactions between more than 20,000 proteins.

To construct a protein interaction network, we download an unweighted version of human protein interactions data from the

BioGRID database and extract required fields such as entrez gene ID of first interactor and entrez gene ID of second interactor from the data to construct a human protein interaction network. An entrez gene ID of interactor is the unique identification number of the gene which is transcripted and translated to the interactor protein. This unique identification number is given and managed by Entrez Gene database at the National Center for Biotechnology Information.

As a result, constructed protein interaction network is represented as an undirected graph $G(V,E)$. In this undirected graph, each vertex v in the vertex set V represents a protein in the BioGRID database and each edge $e=(v,w)$ in the edge set E represents an interaction between the proteins represented by v and w, respectively. So, in the graph $G(V,E)$, two vertices are connected by an edge if and only if there is a reported interaction between the proteins represented by these vertices in the BioGRID database.

Topological terms of protein interactions network

Let $G(V,E)$ is the protein interaction network as previously described. We call a vertex triple (x,y,z), where x,y,z are all in V, and all different to each other, a triangle in protein interaction network, if $(x,y),(y,z),(z,x)$ are edges in the protein interaction

network, i.e. $(x,y),(y,z),(z,x) \in E$. For an edge $e \in E$ of the protein interaction network, we say that a triangle (x,y,z) contains e if $e \in \{(x,y),(y,z),(z,x)\}$.

For a protein interaction network $G(V,E)$, we say vertices v,w in V are adjacent if (v,w) is an edge in this protein interaction network, i.e. $(v,w) \in E$.

For a protein interaction network $G(V,E)$ and an edge $e = (v,w)$, we call v and w as the ends of e.

For a protein interaction network $G(V,E)$, we say a vertex $v \in V$ is incident to an edge e, if v is one of the ends of e.

Weighted protein interaction network

The Weighted protein interaction network is the protein interaction network with an assigned weight at each of its edge. Researchers have come up with the many measurement methods that measures the reliability of given protein interaction and weight each edge of a protein interaction network with this measure. These kind of weighted interaction network is constructed to overcome or identify the false positive interactions generated by high-throughput experiments. One of the very comprehensive weighted protein interaction network is provided by STRING database and figure 2-2 shows an example of a weighted protein interaction network provided by STRING database [Szklarczyk et al. (2011)].

In the figure 2−2, colored circles are the nodes of the weighted protein interaction network and each node represents a protein. Each word besides each circle is the gene symbol of the protein represented by the node. Each blue line connecting two circles is an edge of the weighted protein interaction network and each edge represents a protein interaction between the nodes connected by this edge. The thickness and the intensity of blue color of each edge represents a confidence score given by STRING database and STRING database measures a reliability of a protein interaction with this confidence score.

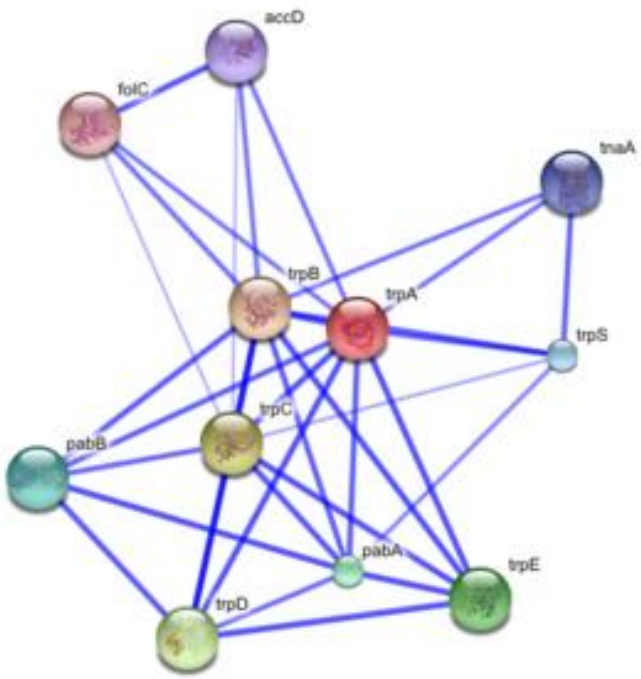

Figure 2−2. Weighted protein interaction network from STRING database

We construct a weighted protein interaction network from a protein interaction network by assigning weight to each edge in the network. So that the constructed weighted protein interaction network is represented as a weighted undirected graph $G(V,E,W)$, where V and E are defined in exactly the same way as in a protein interaction network and W is a weight function $W: E \rightarrow R$, defined by $W(e)$ representing the number of triangles in a protein interaction network which contains e for each edge e in E. We call $W(e)$ the triangle score of e for each edge e in E.

Protein sequence similarity

Sequence similarity searching is the most widely used, and most reliable, strategy for characterizing newly determined sequences [Pearson (2013)]. This characterization means that we can identify proteins or genes with common ancestry through sequence similarity searching. Proteins or genes with common ancestry are called "homologous" and identifying these homologous proteins or genes is important to understand the evolution of proteins or the evolution of protein interaction networks or the evolution of the whole organisms.

Proteins or genes with the excess sequence similarity are considered to be homologs, homologous proteins or genes. In this

context, excess means that given proteins or genes have more sequence similarity than would be expected by chance. Although we can safely assume that proteins or genes with the excess sequence similarity have common evolutionary ancestor, we cannot assume that proteins or genes with insignificant sequence similarity are not homologous since there are many homologous proteins or genes that have insignificant sequence similarities.

In this paper, we defined protein sequence similarity between two proteins as the similarity between nucleotide sequences of the parent genes of these two proteins. A parent gene of a protein is the gene with DNA sequence that is translated to form the protein.

The methods of the evolutionary model of protein interaction networks have direct representations in the protein sequence change. Gene duplications are represented as a production of a copied protein sequences. Subfunctionalization is represented as a divergence of two identical protein sequences resulting in two different protein sequences. Neofunctionalization is represented as random changes of existing protein sequences. Since the methods of the evolutionary model of protein interaction networks are also related to the triangle score, all three methods produce measurable changes in the number of triangles [Gibson et al. (2009)], we make

comparison between the triangle score and the measurement of sequence similarity.

To measure protein sequence similarity, we use BLAST package provided by NIH. By inquiring BLAST package with proteins of interest, we get raw scores and bit scores of the protein sequence similarity. Raw scores are given by the well-known BLAST alignment algorithm described by Altschul et al. [Altschul et al. (1990)] and bit scores are just the log-scaled version of raw scores defined as $[\lambda S - ln(K)]/ln(2)$, where S is a raw score and λ, K are the parameters depending on the context of the algorithm.

Self-interacting proteins

Self-interacting proteins are the proteins which can interact with one or more copies of it. Self-interacting proteins play important roles in cellular functions [Liu et al. (2013)]. Self-interacting proteins have the ability to form homomultimers, which are the macromolecular complex formed by two or more identical proteins. Brenda enzyme database reported that out of total 452 human enzymes for which the subunit composition is listed, a third are monomers. Of the remaining 311 multimers, 199 enzymes form homomultimers [Marianayagam et al. (2004)]. This tells us that self-interacting proteins plays important role in functional regulation in cells. Many multi-protein complexes also formed by

homodimers, that are homomultimers formed with two identical proteins, and some examples of these multi-proteins are proteasome, ribosome, and nucleosome [Ispolatov et al. (2005)]. Protein self-interactions are crucial during cellular signal transduction [Liu et al. (2013)]. This claim is supported by the study showing channel proteins, which control the transport of molecules and ions across cell membranes, relies heavily on their homo-oligomers, oligomers formed by identical proteins [Marianayagam et al. (2004)]. It is also reported that homo-oligomerization can benefit the synthesis of macromolecular complexes by enabling large structure without increasing genome size and increasing error control during synthesis [Liu et al. (2013)].

Other than importance of self-interacting proteins in cellular structures and functions, it is shown that homodimers have twice as many interaction partners than non-self-interacting proteins [Ispolatov et al. (2005)]. So, self-interacting proteins have some relationship with a topological property of protein interaction network.

In a protein interaction network, self-interacting proteins are represented as a vertex with a loop edge, i.e. an edge that connects a vertex to itself. Since self-interacting proteins are believed to

play important roles in cellular functions and the evolution of protein interaction networks, there have been many attempt to predict self-interaction proteins through computational or bioinformatic methods. The reason researchers try to come up with computational or bioinformatic methods for predicting self-interacting proteins is that experimental methods have limited ability to detect self-interactions due to biological artifacts and design limitations [Liu et al. (2013)]. Liu et al. suggested a method to predict self-interacting proteins by considering multiple properties. Considered properties are domain number, evolutionary rate, protein age, degree of a protein in a protein interaction network and betweenness centrality of a protein on a protein interaction network [Liu et al. (2013)]. Furthermore, based on the prediction result they developed the web service called SLIPPER to provide researchers with probabilities of the proteins of interest being self-interacting proteins. An et al. suggested the prediction method by developing a novel feature representation scheme for the amino acid sequence based on LBP(Local Binary Pattern) and employing RVM(Relevance Vector Machine) classifier to the scheme [An et al. (2016)].

Gene duplication and the number of triangles

For a given protein interaction network $G(V,E)$, let us consider a duplication of a gene g and let p be a protein that is produced by translating g. Let $N[v]$ be the set of vertices which are adjacent to v and define $N(v)$ as the set $N[v]-\{v\}$.

The duplication of gene g is represented in $G(V,E)$ by adding a vertex w to V and adding (w,x) to E for each x in $N[v]$. Now, let t be the number of edges with both ends in $N(v)$ and d be the size of $N(v)$. If the protein p is a self-interacting protein then the duplication of g produces $t+d$ triangles. For example, in figure 2-3, the protein A with degree 4 is duplicated and A is a self-interacting protein. After duplication, network has 1+4=5 more triangles.

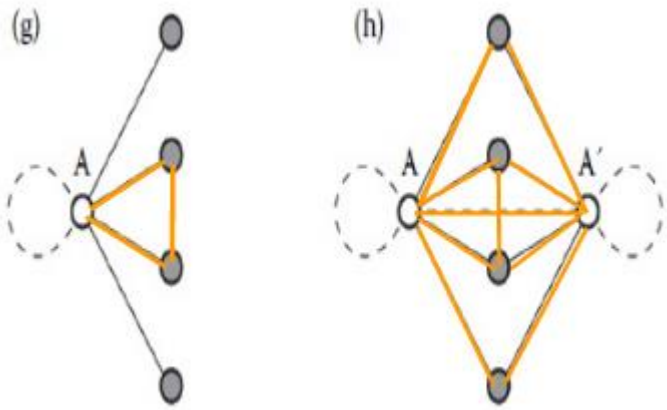

Figure 2-3. Duplication of self-interacting protein [Vazquez (2010)]

Otherwise, i.e. p is not a self-interacting protein, the duplication of g produces t triangles. For example, in figure 2-4, the protein A with degree 4 is duplicated and A is not a self-interacting protein. After duplication, network has one more triangle.

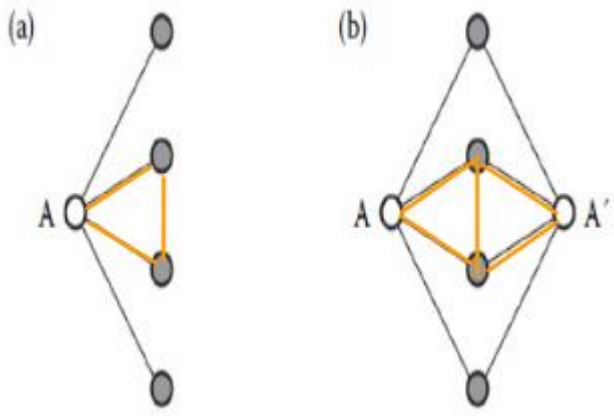

Figure 2-4. Duplication of non-self-interacting protein [Vazquez (2010)]

Based on this calculation, we come up with a conjecture that a vertex corresponding to a self-interacting protein is more likely to incident to an edge with high triangle score. To prove this conjecture, we make comparison between the triangle scores of self-interacting proteins and the triangle scores of non-self-interacting proteins.

Construction of known self-interacting protein list

To get a currently known self-interacting protein list, firstly, we download the whole human protein data from the NIH Gene database. After downloading the whole data, we extract interacting partners of each human protein and check whether a protein has itself as an interacting partner. If a protein has itself as an interacting partner, then we add this protein to the list we are constructing.

Phylogenetic age

Phylogenetic age is defined through evolutionary relationships among species and the study of these evolutionary relationships among species is the one of the main areas of phylogenetics. The result of phylogenetic study about this area is a hypothesis about the evolutionary relationship among species and these relationships are commonly represented as a branching diagram called a species tree. An example of a species tree is shown in figure 2-5.

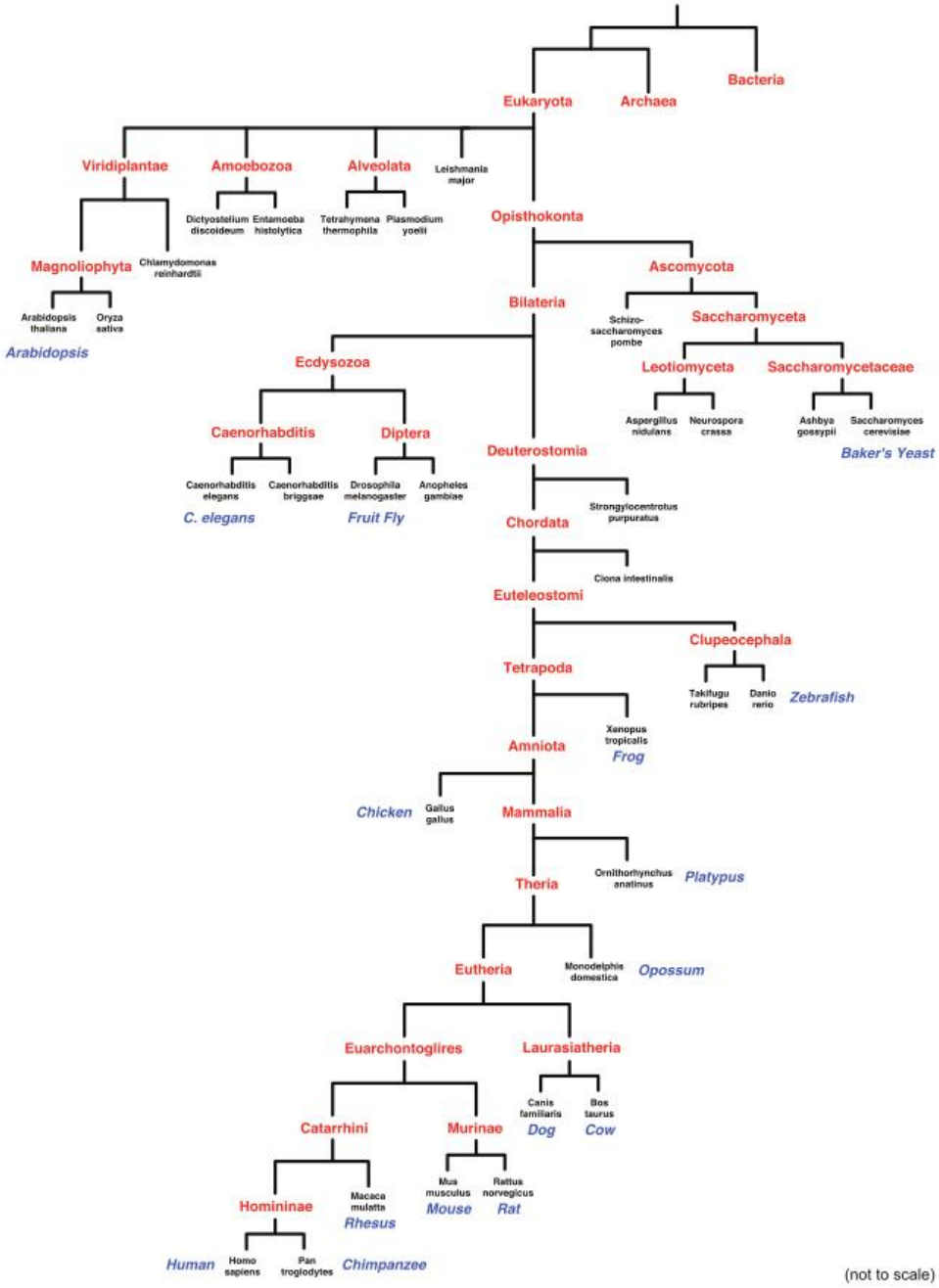

Figure 2-5. Species tree

In the species tree, each node represents an organism and a node with descendants represents the most recent common ancestor of the descendant nodes.

Phylogenetic studies go through iterative steps to reconstruct and update evolutionary histories. At first, researchers observe the nucleotide or protein sequences and infer evolutionary events by combining the observations and their sequence evolution model. With these inferred evolutionary events, researchers reconstruct or update evolutionary histories. During these steps, researchers may obtain more information regarding sequence evolution or evolutionary events and this leads to refined sequence evolution model.

Phylogenetic age is represented as an organism in the evolutionary history. If a protein p has a phylogenetic age of an organism O, then this protein p is most likely diverges from the parent organism of the organism O on the species tree.

ProteinHistorian

ProteinHistorian is an integrated web platform which provides easy to use estimation tools for the phylogenetic age [Capra et al. (2012)]. ProteinHistorian computes the protein age, i.e. the phylogenetic age of a protein, based on a species tree, a

protein family database, and an ancestral family reconstruction algorithm.

Species trees: A species tree is a branching diagram which explains the inferred evolutionary relationships among species. ProteinHistorian uses a species tree based on the modified version of NCBI taxonomy database to reflect recent research. For each internal node, i.e. a node with descendants, divergence time is estimated from TimeTree database.

Protein family databases: Protein family is a group of proteins that are related in the evolutionary term. Proteins in the same protein family are descended from the common ancestor. A protein family database holds the information about protein families and provides protein family predictions for the given proteins.

Ancestral history reconstruction algorithms: For each protein family, an ancestral history reconstruction algorithm reconstructs evolutionary history based on its assumption. ProteinHistorian provides two reconstruction algorithms, Dollo parsimony and Asymmetric Wagner parsimony. Dollo parsimony assumes that losing a complex structure is much more common event than gaining one. Thus Dollo parsimony assumes a single gain for each protein family followed by many losses. Under Dollo parsimony an origin of family is the most recent common ancestor of all species in

observed data. Wagner parsimony allows multiple gain and loss events. When using Wagner parsimony, one can set weights on the relative likelihood of gaining and losing.

So, ages of a protein estimated by ProteinHistorian is the time at which recognizable homology of the protein first evolved. [Capra et al. (2012)]

Estimation of phylogenetic age

We choose the edge $e = (v, w)$ with highest triangle score in the constructed weighted protein interaction network and collect proteins represented by v and w. Then delete the edge e from the network and choose the edge with highest triangle score in this modified network and collect the proteins incident to the edge. We repeat this procedure until we collect 300 proteins. Use the collection as an input to *ProteinHistorian,* we get the phylogenetic age of proteins participating in an interaction corresponding to an edge with high triangle score.

DisGeNet

DisGeNet is a web platform developed by Pinero et al. to help researchers in discovering the genetic properties of human diseases [Pinero et al. (2015)]. This kind of easily accessible and comprehensive data base has been needed by many researchers since there has been vast increase of biomedical data in public

sources. Researchers have attempted to utilize and integrate these genomic, phenomic, and environmental information for the better understanding of disease mechanisms and the implementation of personalized medicine [Sarkar et al. (2011); Topol et al. (2014); Pinero et al. (2015)]. These pursuitments will make researchers and clinical practitioners to increasingly rely on the data of the genetic determinants of disease and the availability of comprehensive knowledge sources on disease genes and tools should lay the basis for the pursuitments [Pinero et al. (2015)]. DisGeNet aims at aiding the researchers and clinical practitioners in these pursuits and more clinical actions that can be actualized through biomedical data.

DisGeNet is one of the largest repositories relating genes and diseases and it is currently holding 380,000 associations between more than 16,000 genes and 13,000 diseases. DisGeNet integrates text-mined data, information on Mendelian and complex diseases, and data from animal disease models. DisGeNet provides a score of an association between a disease and a gene based on the supporting evidence to prioritize gene-disease models. The score accounts for the number of sources that report the association, the type of curation of each of these sources, the animal models where

the association has been studied, and the number of supporting publications from text-mining based sources [Pinero et al. (2015)].

Disease selection

We make an inquiry to DisGeNet database about 70 complex diseases chosen by Ghiassian et al. in their research [Ghiassian et al. (2015)]. This inquiry gives us a list of disease associated proteins for each of 70 complex diseases and we call this list the known set of associated proteins for a disease. Though DisGenet provides scores for each associated protein, we do not use any threshold score to filter out proteins. As a result, the known set of associated proteins for a disease contains every protein that have at least one supporting evidence associating the protein and the disease of interest. Among 70 complex diseases, we choose 45 diseases that have at least 20 associated proteins. For each of these 45 diseases, we make the prediction of associated proteins based on triangle scores.

Prediction of disease associated protein

We define disease module as a collection of disease associated proteins for the disease of interest. Ghiassian et al. showed that disease modules have connectivity properties different from other modules such as functional modules. Motivated by their work on disease module prediction, we modified the disease module

prediction algorithm, suggested by Ghiassian et al., based on triangle scores.

Our disease module prediction algorithm:

Firstly, we randomly choose 10 proteins from the known set of associated proteins of a disease and call this collection of chosen proteins as seed protein set. On the weighted protein interaction network, select all edges each of which has one end representing an element of the seed protein set and the other end representing a protein not in the seed protein set. If a selected edge is (v, w) then either v is in seed protein set and w is not in seed protein set or v is not in a seed protein set and w is in a seed protein set. From selected edges, choose the one with highest triangle score. Between the two proteins incident to the chosen edge, choose the one that is not in the current seed protein set and add it to the seed protein set to expand the current seed protein set. We repeat these steps, except the first step, with newly obtained seed protein set until the size of a seed protein set reaches 110. After we get a seed protein set of size 110, we construct a protein set of size 100 by deleting 10 original seed proteins from the seed protein set of size 110 which we have gotten at the end of the iteration. We call this protein set of size 100 as a set of predicted proteins.

Evaluation of the prediction

Since the full set of associated proteins for a disease is unknown, we assume that a predicted protein is associated with a disease of interest if it is participating in at least one of significantly enriched pathways of the known set of associated proteins for a disease. The known set of associated proteins for a disease is obtained through inquiries to the DisGeNet, as previously mentioned. To obtain the significantly enriched pathways for a disease, we gather every pathway from REACTOME pathways database that has at least one protein in the known set of associated proteins as its participant. With collected pathways, we calculate the p−value to measure the significance with significance level 0.05. Under this assumption, we compute a precision of the prediction for each of 45 diseases.

Human Protein Reference Database

To compare our algorithm of inferring gene phenotype relationship with the algorithm suggested by Li et al. [Li et al. (2010)], we used the same protein interaction network as the previous work when we construct a heterogeneous network. Li et al. derived a protein interaction network from Human Protein Reference Database (HPRD) [Peri et al. (2003)]. So, we downloaded the same version of HPRD as the Li et al.

HPRD contains the data of curated proteomic information pertaining to human proteins and all the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyze the published data [Peri et al. (2003)].

OMIM (Online Mendelian Inheritance in Man) database

OMIM is electronic counterpart of Mendelian Inheritance in Man (MIM), which has been published in 12 print editions since 1966 and is a compendium of information on genetic disorder and genes. Curation and the editorial decisions concerning the database is made at Johns Hopkins University School of Medicine. Distribution of OMIM and software development are provided by the National Center for Biotechnology Information at the National Library of Medicine [Hamosh et al. (2005)].

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM contains the data about relationships between mendelian disorders and over 15,000 genes. OMIM focuses on the relationship between phenotype and genotype [Hamosh et al. (2005)].

MimMiner

Even though there is a detectable linkage between phenotype clusters and the function of the underlying genes, human disease phenotype data sets, that exist before MimMiner, such as OMIM are not fit for the purpose of systematic analysis of phenotypes [van Driel et al. (2006)]. MimMiner aims at aiding researchers in systematic analysis of phenotypes. Moreover, MimMiner provides similarity scores between phenotypes so that researchers can compare genes with known phenotypes [van Driel et al. (2006)]. The similarity scores between phenotypes are calculated by analysing OMIM database with various text mining algorithms and MimMiner provides fast and easy searching method on their web page [van Driel et al. (2006)].

Phenotype

A phenotype is a description of physical characteristics or traits of an organism. A phenotype results from the expression of a genotype of an organism. For example, let us assume that whether petal color of pea plants is controlled by a single gene. Furthermore, assume that among the alleles in the gene, only two alleles are commonly found. Now, denote these two alleles as B and b, respectively. Then there are three genotypes: BB, Bb, bb. Each of these genotypes has phenotype, i.e. petal color, related to it. Let us assume the phenotypes are assigned to genotypes as follows:

$BB(purple), Bb(purple), bb(white)$. In this case, only one B allele in the genotype make petal color of a pea plant purple and this allele B is called dominant allele. To petal color of a pea plant be white, the plant needs the genotype entirely made up of allele b and this allele b is called recessive allele.

We define a phenotype entry as an MIM record, same as previous works [van Driel et al. (2006); Wu et al. (2008); Li et al. (2010)].

Weighted gene network

The gene network is just a protein interaction network obtained from Human Protein Reference Database. We weight each edge of the gene network with its triangle score to construct the weighted gene network.

Phenotype network

In this paper, the phenotype network is represented as a weighted undirected graph $G(P, E, W)$. In this graph, each vertex p in P represents a phenotype entry, which is obtained from OMIM database, and for $p, q \in P$, $(p, q) \in E$ if q is one of five most similar phenotype entries of p. Similarity score is calculated by inquiring each phenotype entry on MimMiner web page and top five high scored phenotype entries are assumed to be the five most similar

phenotype entries. W is the weight function that weights each edge with the similarity score given by MimMiner.

Gene−phenotype network

The gene−phenotype network is represented as a bipartite graph $G(V,P,E)$. In this graph, each vertex $v \in V$ represents a protein and each vertex $p \in P$ represents a phenotype entry. For $v \in V$ and $p \in P$, $(v,p) \in E$ if the protein represented by v and the phenotype entry represented by p have a known relationship. This known relationship between gene and phenotype is obtained from OMIM database.

Heterogeneous network

The heterogeneous network is constructed by connecting the weighted gene network and the phenotype network using the gene−phenotype network. So, the heterogeneous network is represented as a weighted undirected graph $G(V,E,W)$. In this graph, V is the union of the vertex sets of the weighted gene network and the phenotype network and E is the union of the edge sets of the weighted gene network, the phenotype network and the gene−phenotype network. The weight function $W: E \rightarrow R$ is defined as follows:

$W(e) = the\ \ triangle\ \ score\ \ of\ \ e$, if both ends of e represent genes

$W(e) = the\ similarity\ score\ of\ e$, if both ends of e represent phenotype entries

$W(e) = 1$, if one end of e represents gene and another end represents phenotype entry

Random walk with restart

Random walk with restart is a ranking algorithm [Kohler et al. (2008)]. In this algorithm, a random walker starts from a seed node or a set of seed nodes on a given network and moves to one of adjacent vertices of a seed node randomly at each step. Finally, all the nodes in the network are ranked by the probability of the random walker reaching this node [Li et al. (2010)]. Random walk with restart is simulated through multiplications of a transition matrix and a probability vector. The transition matrix for the given network M is a $\mathrm{n} \times \mathrm{n}$ matrix where n is the number of nodes in the network and $i,j-$th term M_{ij} is the transition probability from node i to node j. The probability vector at step s is denoted as $\mathbf{p_s}$ and $i-$th term of $\mathbf{p_s}$ is the probability of finding random walker at node i at step s. Let $\mathbf{p_0}$ be the initial probability vector then $\mathbf{p_{s+1}}$ is given by

$$\mathbf{p_{s+1}} = (1-\gamma)M^T\mathbf{p_s} + \gamma\mathbf{p_0}$$

where $\gamma \in (0,1)$ is the restart probability, which means that at each step, random walker can return to seed nodes with the

probability γ. We calculate the probability vector at each step until the difference between $\mathbf{p_{s+1}}$ and $\mathbf{p_s}$ is less than 10^{-10}, where the difference is measured by $L1$ norm.

For our heterogeneous network, the transition matrix M can be written as

$$M = \begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_P \end{bmatrix}$$

where M_G and M_P are the transition matrix for the weighted gene network and the phenotype network, respectively. M_{GP} and M_{PG} are the transition matrix for the gene−phenotype network. Exact terms are defined as follows:

$$(M_{GP})_{ij} = \begin{cases} \lambda B_{ij} / \sum_j B_{ij}, & \text{if } \sum_j B_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$(M_{PG})_{ij} = \begin{cases} \lambda B_{ji} / \sum_j B_{ji}, & \text{if } \sum_j B_{ji} \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

where B is the adjacency matrix for the gene−phenotype network and is the jumping probability, that is the probability of the random walker jumping from the gene network to phenotype network or vice versa. Let A_G and A_P be the adjacency matrices for the weighted gene network and the phenotype network, respectively, so each term of A_G is a triangle score and each term of A_P is a

phenotype similarity score, we can define exact terms of M_G and M_P as follows:

$$(M_G)_{ij} = \begin{cases} (A_G)_{ij} / \sum_j (A_G)_{ij}, & \text{if } \sum_j B_{ij} = 0 \\ (1-\lambda)(A_G)_{ij} / \sum_j (A_G)_{ij}, & \text{otherwise} \end{cases}$$

$$(M_P)_{ij} = \begin{cases} (A_P)_{ij} / \sum_j (A_P)_{ij}, & \text{if } \sum_j B_{ji} = 0 \\ (1-\lambda)(A_P)_{ij} / \sum_j (A_P)_{ij}, & \text{otherwise} \end{cases}$$

The initial probability vector for our heterogeneous network is defined with the initial probability vector of the weighted gene network, $\mathbf{u_0}$, and the initial probability vector of the phenotype network, $\mathbf{v_0}$, and the parameter $\eta \in (0,1)$. The exact terms are defined as follows:

$$(\mathbf{u_0})_k = 1/n_g$$
$$(\mathbf{v_0})_l = 1/n_p$$

where ng is the number of seed nodes in the weighted gene network and np is the number of seed nodes in the phenotype network. Finally the initial probability vector for our heterogeneous network $\mathbf{p_0}$ is defined as:

$$\mathbf{p_0} = \begin{bmatrix} (1-\eta)\mathbf{u_0} \\ \eta \mathbf{v_0} \end{bmatrix}$$

제 3 장

Results

Comparison between triangle score and existing curated score

Before go deep into the evolutionary studies of protein interaction networks, we make comparison between triangle score and existing curated score to get some insight to the protein interaction network with triangle score. Our comparison is made between the triangle score, that we defined in this paper, and the existing curated score provided by STRING database. STRING database is constructed by Szklarczyk et al. [Szklarczyk et al. (2011)]. STRING database provides comprehensive coverage of protein interactions since it combines protein interactions from many other databases such as BioGRID and IntAct. More importantly, STRING database provides confidence score of protein interactions in the database and this confidence score is based on genomic context, high-throughput lab experiments and coexpression studies. By downloading protein interaction data from STRING database and treating provided confidence scores as edge weights, we get another weighted protein interaction network based on the confidence scores from STRING database. We call this

weighted protein interaction network as the protein interaction network based on STRING score and, naturally, call weights of this protein interaction network as STRING scores. We make a comparison between the weighted protein interaction network based on triangle score and the weighted protein interaction network based on STRING score.

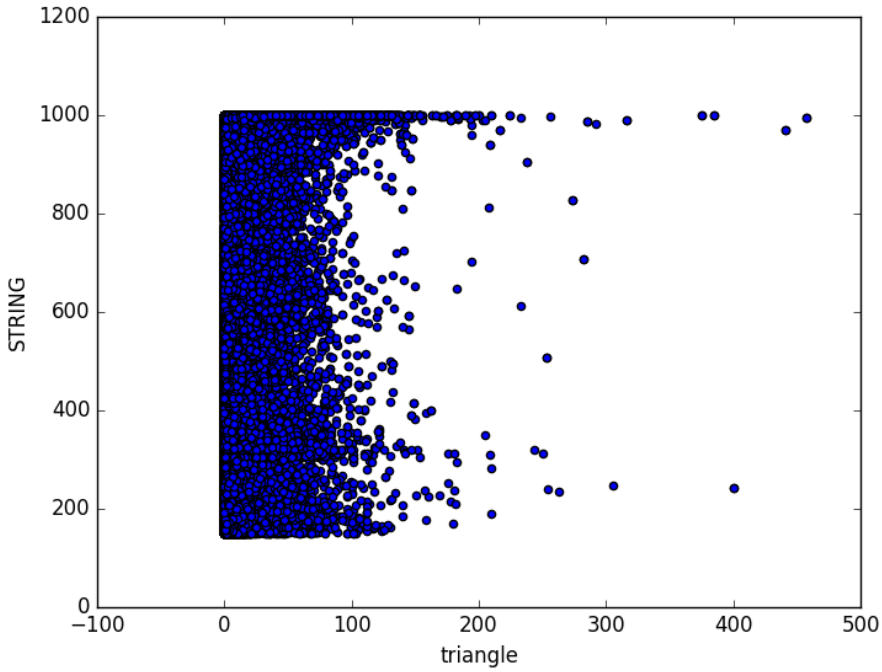

Figure 3−1. Scatter plot of a comparison between the triangle score and STRING score.

Figure 3−1 is the scatter plot resulted from the comparison between triangle score and STRING score. In the scatter plot, $x-$axis represents triangle score and $y-$axis represents STRING

score and each blue dot represent an edge of the protein interaction network. The result tells us that if a protein interaction has triangle score greater than 300 then this interaction is likely to have a highest STRING score. However, as clearly shown in the plot, for the interactions with triangle score that is not significantly high, we cannot say anything about its STRING score. The Pearson correlation coefficient is 0.189.

Protein sequence similarity and triangle score

We extract the whole protein list from the protein interaction network and make inquiry to the BLAST package to get bit scores of the proteins. Since inquiring the BLAST package with a protein list give us bit scores between every pair of proteins in the list, we filter the result to get the bit scores between proteins that are adjacent in the protein interaction network. After this procedure, we make a comparison between triangle scores and filtered bit scores.

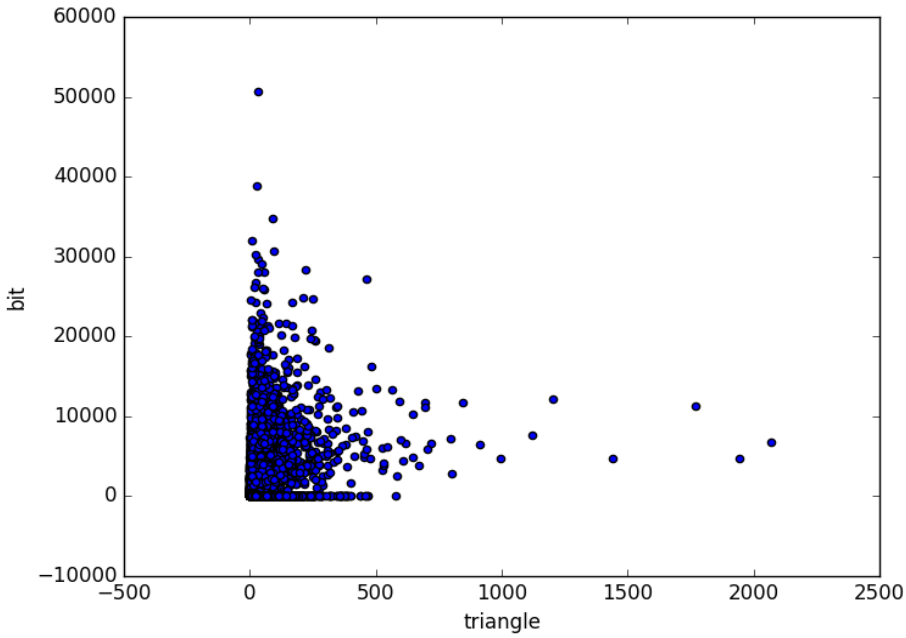

Figure 3−2. Scatter plot of a comparison between triangle score and bit score.

In figure 3−2, each blue dot represents an edge of the protein interaction network, x−axis represents the triangle score, and y−axis represents the bit score.

The result shows that interactions with high triangle scores, i.e. triangle score greater than 500, have low bit scores, i.e. less than 13000. However, all of these interactions, except one interaction, have the same parent gene for participating proteins. This result is due to the fact that participants, i.e. participating proteins, of interactions with high triangle score tend to have parent

genes with low nucleotide lengths. For the interactions with high bit scores, i.e. bit score greater than 20000, triangle score is low, i.e. less than 300, and its participants have parent genes with long nucleotide sequences, greater than 10000. However, this clear trend is existing for just small proportion of protein interactions of interest and generally there is no relationship exists between the bit score and the triangle score. The Pearson correlation coefficient is 0.223.

Relationship between self-interacting protein and triangle score

In an effort to explore the properties of triangle score, we draw following plots in which relationship between self-interacting proteins and triangle scores is expressed.

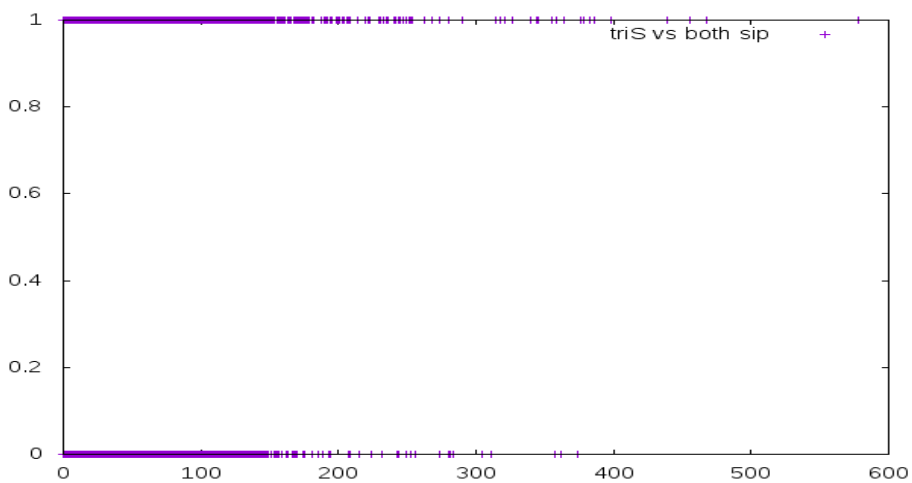

Figure 3-3. Relationship between triangle scores and self-interacting proteins

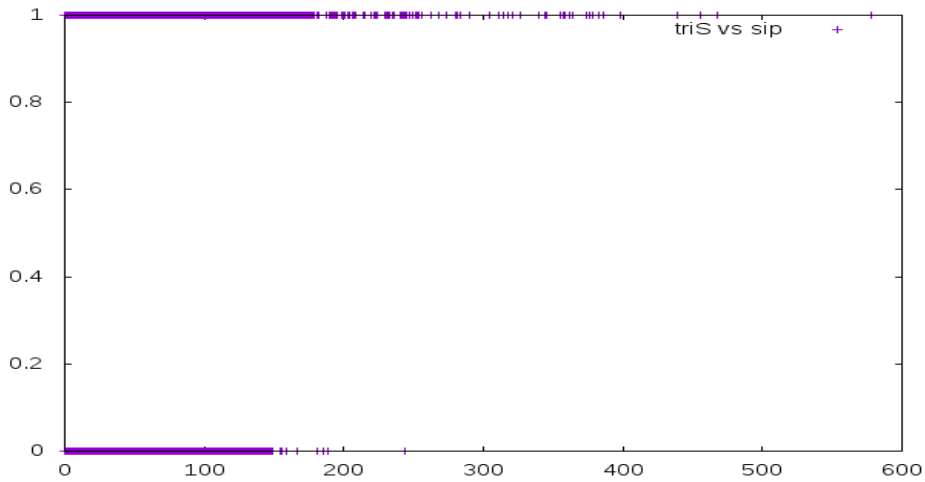

Figure 3-4. Relationship between triangle scores and self-interacting proteins

In figure 3-3, each cross represents an edge of the protein interaction network, x-axis represents triangle scores of edges and y-axis represents whether or not both ends of an edge are self-interacting proteins. If both ends are self-interacting proteins, then the value of y-axis is one and otherwise zero.

In figure 3-4, x-axis represents triangle scores of edges and y-axis represents whether or not at least one of the ends of an edge is self-interacting protein. If at least one of the ends is self-interacting protein, then the value of y-axis is one and otherwise zero.

As clearly shown in the plots, if an edge has significantly high triangle scores, i.e. greater than 300, then we can safely

assume at least one of the ends of this edge is a self-interacting proteins. However, if an edge has a triangle score that is not significantly high then we cannot say anything about the ends of given edge being self-interacting proteins. For both ends of an edge to be self-interacting proteins, we need an edge to have triangle score greater than 400.

Phylogenetic age of proteins with high triangle score

We define the triangle score of a protein as the highest triangle score among the triangle scores of edges which are incident to it. Since the evolutionary models of protein interaction network tell us that existence of just one interaction with high triangle edge incidents to a protein indicates that this protein is old in phylogenetic age. Moreover, using the sum or the mean of triangle scores of the edges incident to a given protein to estimate phylogenetic age of proteins gave us worse result than using the maximum of triangle scores of edges incident to a given protein. Thus our definition of the triangle score of a protein is justified by the theory and the practice. We extract proteins with top 300 triangle scores and make an estimation of phylogenetic age for these proteins.

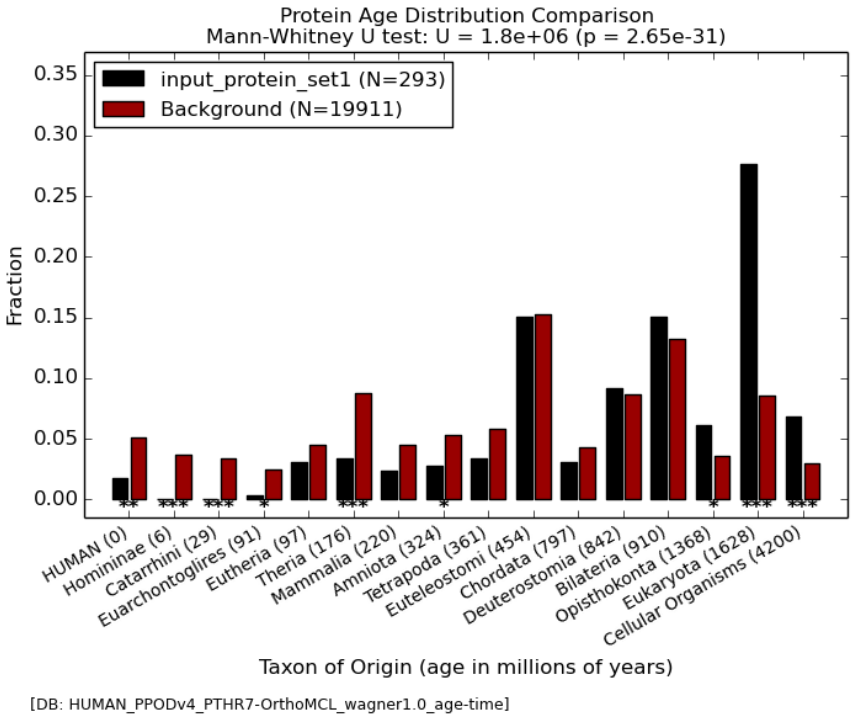

Figure 3-5. Estimation of phylogenetic age of proteins with high triangle score.

ProteinHistorian gives us a result as a bar graph in which y-axis represents a fraction of certain aged proteins in the given set and x-axis represents categories of the phylogenetic age. In the given chart, red bars represent the result of phylogenetic age estimation on the whole human proteins and black bars represent the result of phylogenetic age estimation on the set of proteins with top 300 triangle score. Though our original input size was 300, the result is accounted for only 293 proteins because *ProteinHistorian*

failed to estimate the phylogenetic age for some proteins in our input.

The result shows that the set of proteins with high triangle score has a significantly high number of old proteins in terms of phylogenetic age. More specifically, while the average age of whole human proteins is only 681.4, the average age of our input, i.e. proteins with high triangle score, is 1163.5. Medians of the result, also, shows the clear difference; the median for whole human proteins is 454.6 and the median for our input is 910.0.

Analyzing result for each category of the phylogenetic age reveals the same trait that proteins with high triangle score are significantly older than whole proteins. The oldest category of phylogenetic age is Cellular organisms, which indicates the first appearance of cellular organisms in evolution, and the fraction of this category is significantly higher in our input than in the whole human proteins with a significance level 0.001. The second oldest category of phylogenetic age is Eukaryota and the fraction is, also, significantly higher in our input with a significance level 0.001. For the categories of young phylogenetic age, the comparison result is reversed. The youngest category of phylogenetic age is Human and the fraction of this category is significantly lower in our input than in the whole human proteins with a significance level 0.01. The

second and third youngest categories are Homininae and Catarrhini and the fraction is significantly lower for both categories in our input with a significance level 0.001. For the remaining categories except for Theria, difference of fractions between our input and whole human proteins is not significant or negligible.

Precision of the prediction of disease associated proteins

We define precision of the prediction for a disease as a ratio of the number of correctly predicted proteins to the whole number of predicted proteins. We assume that a predicted protein is a correct prediction if this protein is participating in at least one of significantly enriched pathways of the disease of interest. Under this assumption, we make predictions about disease associated proteins for 45 complex diseases and calculate precisions.

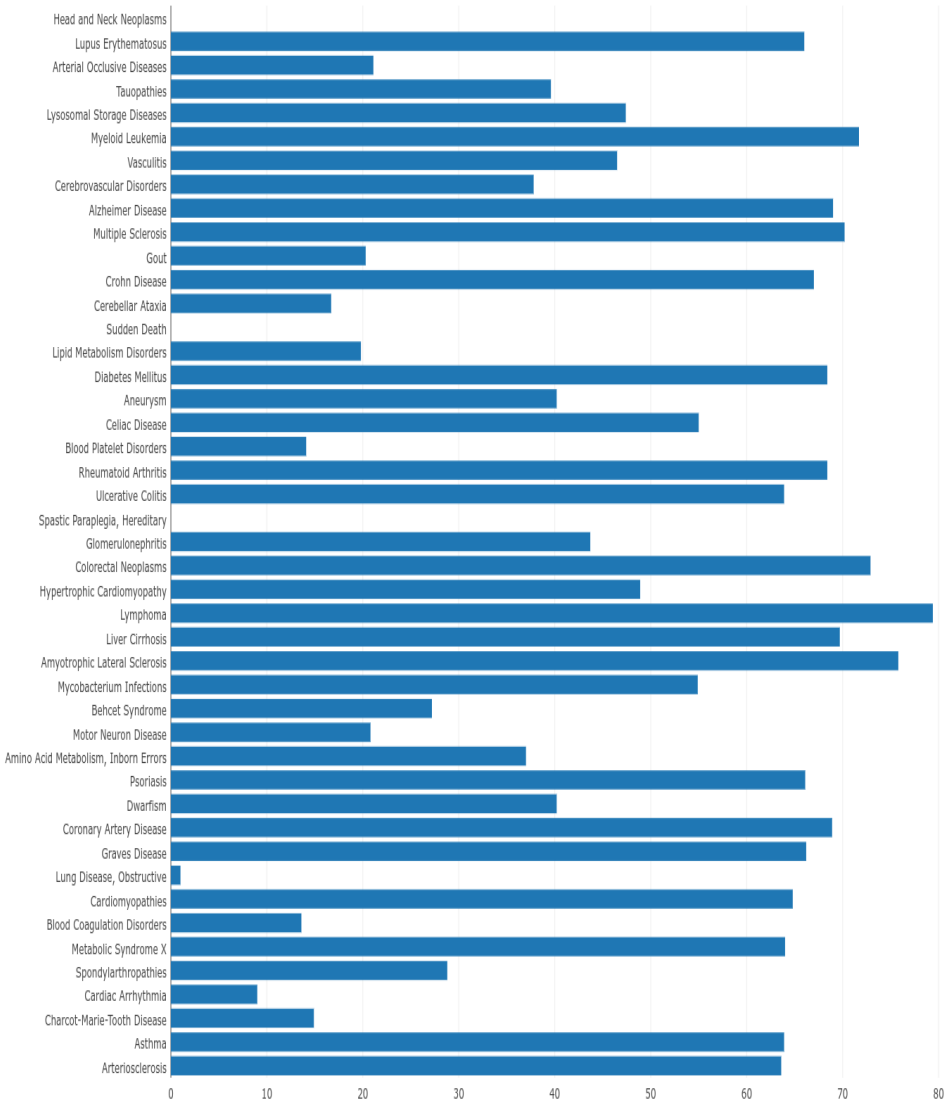

Figure 3−6. Precisions of the predictions for each of 45 diseases

The result of calculation is shown in figure 3−6 as a bar graph with y−axis represents names of diseases and x−axis represents the precision. In the bar graph, a precision is an average of precisions from ten predictions with randomly assigned seed

proteins. The precisions of predictions vastly differ between diseases. The disease that marks the highest precision with our prediction method is lymphoma and the precision is 79.4. Our method predicts disease associated proteins with precision greater than 70 for five diseases: Amyotrophic lateral sclerosis, colorectal neoplasms, lymphoma, myeloid leukemia, multiple sclerosis. Our method also predicts disease associated proteins for five diseases with fairly good precision, i.e. less than 70 and greater than 68: Alzheimer' s disease, coronary artery disease, diabetes mellitus, liver cirrhosis, rheumatoid arthritis. However, our method performs poorly in predictions of disease associated proteins for aneurysm, arterial occlusive diseases, Behcet syndrome, blood coagulation disorders, blood platelet disorders, cerebellar ataxia, Charcot-Marie-Tooth disease, dwarfism, glomerulonephritis, obstructive lung disease, spastic paraplegia, spondyloarthropathies, sudden death, tauopathies and cardiac arrhythmia. Overall, our method predicts disease associated proteins for cancer related diseases with a good precision but for other diseases performance is generally poor.

Overlaps of the prediction of disease associated proteins

Since the prediction of disease associated proteins performs well in only a specific kind of diseases, we feel the need to check

whether our method predicts the same proteins regardless of a seed protein sets. We run our prediction method 30 times with randomly chosen seed protein set of size 10 in each iteration. Thus, different from our original method, the seed protein set is not selected from the known associated proteins of the disease of interest. The seed protein set is randomly selected from the whole human proteins. To check whether our prediction method predicts the same proteins, we make heat maps with the results of the iteration.

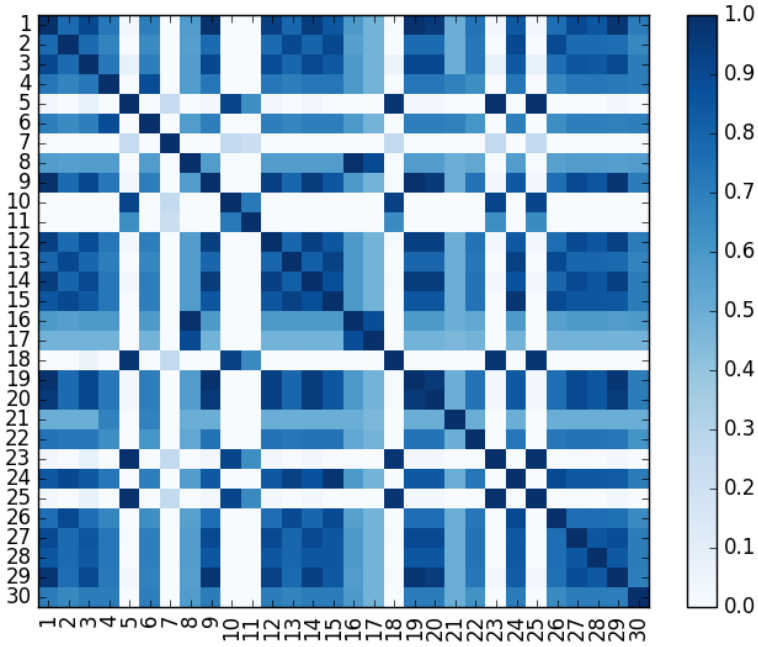

Figure 3-7. Overlaps of the prediction of disease associated proteins with random seed sets

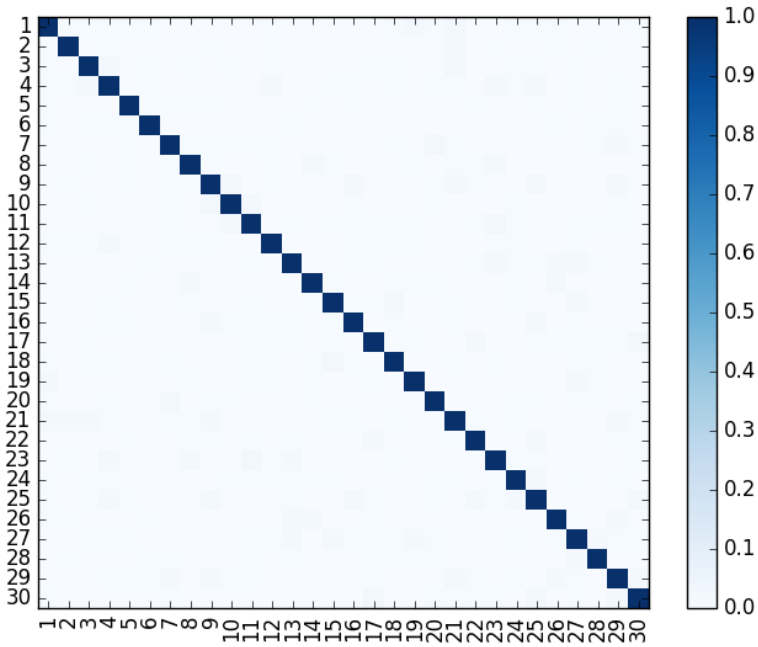

Figure 3-8. Overlaps of randomly chosen 100 proteins

Figure 3-7 is a heat map of the overlaps between predicted proteins via our prediction method. Color of the square in i-th row and j-th column represents a ratio of the number of proteins predicted both from i-th iteration and j-th iteration to the number of proteins predicted by i-th iteration. the scale of color is presented at the right side of the plot with a color bar.

Figure 3-8 is a heat map of the overlaps between randomly selected 100 proteins from the whole human proteins. A comparison of the figure 3-7 and figure 3-8 shows us that our prediction

method highly overlaps between each predictions than randomly expected.

Inferring gene−phenotype relationship

To compare the result with the algorithm suggested by Li et al. we used the same data and the same evaluation measure as the suggested algorithm. The evaluation is made through leave−one−out cross−validation. In each round, we remove one gene−phenotype link (v, p) from the heterogeneous network and set the phenotype p and the genes related to this phenotype as the seed nodes. If the gene v, which was initially left out, is ranked as top 1 after the random walk with restart then we mark this gene as successfully inferred by the algorithm. In this comparison, we set as 0.7 and and as 0.5. The main and only difference between the algorithm suggested by Li et al. and our algorithm is the use of weighted gene network. While Li et al. use unweighted version of the gene network to construct a heterogeneous network, our algorithm use weighted version of the gene network, which is weighted with the triangle score, to construct a heterogeneous network. However, the result of the comparison shows that our algorithm performs worse than the algorithm suggested by Li et al. While Li’ s algorithm infer 254 genes successfully, our algorithm predicts only 252 genes successfully.

제 4 장

Conclusions

We have shown that the triangle score is very different from STRING score since only few edges with significantly high triangle scores are strongly related to STRING score.

The protein sequence similarity is not strongly related to the triangle score and only in few extreme cases the triangle score is related to the protein sequence similarity.

We have failed to find meaningful difference between the triangle scores of self-interacting proteins and the triangle scores of non-self-interacting proteins.

We have shown that if a protein is a participant of an interaction that is included in many triangles in a protein interaction network then the protein is most likely old in terms of phylogenetic age. This result shows, at some level, that considering triangles in protein interaction networks as important topological property is meaningful for evolutionary models. Though we did not make an estimation of phylogenetic age for each category of phylogenetic age, we have shown that there is a relationship between a

topological property, which arises from the evolution of protein interaction networks, and phylogenetic age of proteins.

We have attempted to predict disease associated proteins for complex diseases by utilizing triangle scores and the known associated proteins for each disease. While, our prediction method performs well in predicting disease associated proteins for cancer related diseases such as lymphoma, our method yields poor results for predicting disease associated proteins in general. This result may indicate that cancer related diseases are more likely related to proteins with older phylogenetic age than other diseases.

Overlap study of prediction of disease associated genes has shown that our prediction method favors the proteins which are the participants of the interactions with high triangle score. We have tried to overcome this bias by connecting the protein interaction network with other biological networks, more specifically the phenotype network. However, we have failed to improve the existing algorithm on the heterogeneous network, which is the connected network of the gene network and the phenotype network, with the triangle score and thus have failed to overcome the bias present in our disease associated gene prediction. This suggests that we need to find other biological network that works well with the triangle scored protein interaction network or we have to find a

way to reduce the bias by altering the triangle scored protein interaction network itself.

Bibliography

Altschul,S. et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* 215(3), 403-410.

An,J.Y. et al. (2016). Robust and accurate prediction of protein self-interactions from amino acids sequence using evolutionary information. *Mol. BioSyst.,* 12, 3702-3710.

Barabasi,A.L. et al. (2011). Network medicine: A network-based approach to human disease. *Nat Rev Genet.,* 12(1), 56-68, doi:10.1038/nrg2918.

Breitkreutz,B.J. et al. (2003). The Grid: the general repository for interaction datasets. *Genome Biol.,* 4(3), R23.

Breitkreutz,B.J. et al. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.,* 36(Database issue), D637-40.

Capra, J. et al. (2012). ProteinHistorian: Tools for the comparative analysis of eukaryote protein origin. *PLOS Comput. Biol.,* 8(6), e1002567, doi:10.1371/journal.pcbi.1002567.

Chatr-Aryamontri,A. (2013). The BioGRID Interaction Database: 2013 update. *Nucleic Acids Res.,* 41, 816-823.

Chatr-Aryamontri,A. (2015). The BioGRID Interaction Database: 2015 update. *Nucleic Acids Res.,* 43, 470-478.

Chatr-Aryamontri,A. (2017). The BioGRID Interaction Database: 2017 update. *Nucleic Acids Res.,* 2017(1).

Domazet-Loso,T. et al. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology,* 8(66).

Ghiassian,S. et al. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLOS Comput. Biol.* 11(4), e1004120, doi:10.1371/journal.pcbi.1004120.

Gibson,T. et al. (2009). Questioning the ubiquity of neofunctionalization. *PLoS Comput. Biol.,* 5(1), e1000252, doi:10.1371/journal.pcbi.1000252.

Gibson,T. et al. (2011). Improving evolutionary models of protein interaction networks. *Bioinformatics,* 27(3), 376-382.

Hamosh,A. et al. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.,* 33 (Database Issue), D514-D517.

Ispolatov,I. et al. (2005). Binding properties and evolution of homodimers in protein-protein interaction network. *Nucleic Acids Res.,* 33(11), 3629-3635, doi: 10.1093/nar/gki678.

Kohler,S. et al. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet,* 82, 949-958.

Li,Y. et al. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics,* 26(9), 1219-1224.

Liu,Z. et al. (2013). Proteome-wide prediction of self-interacting proteins based on multiple properties. *Molecular & Cellular Proteomics,* 12, 1689-1700.

Marianayagam,N. (2004). The power of two: protein dimerization in biology. *Cell,* 29(11), 618-625.

Mount,D. (2004). "Sequence Database Searching for Similar Sequences". *Bioinformatics: Sequence and Genome Analysis,* 2nd edition.

Nabieva et al(2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics,* 21(Suppl 1), i302-i310.

Pearson,W. (2013). An introduction to sequence similarity ("Homology") searching. *Curr. Protoc. Bioinformatics,* 2013 Jun, 0 3, 10.1002/0471250953.bi0301s42, doi: 10.1002/04712509053.bi0301s42.

Peri et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.,* 13, 2263-2371.

Peterson,G.J. et al. (2012). Simulated evolution of protein-protein interaction networks with realistic topology. *Plos ONE,* 7(6), e39092, doi:10.1371/journal.pone.0039052.

Pinero,J. et al.(2015). DisGeNet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database,* Vol. 2015, article ID bav028, doi:10.1093/database/bav028.

Sarkar,I.N. et al. (2011). Translational bioinformatics: linking knowledge across biological and clinical realms. *J. Am. Med. Inform. Assoc.,* 18, 354-357.

Sole,R. et al. (2002). A model of large-scale proteome evolution. *Adv. Complex Syst.,* 5, 43-54.

Stark,C. et al. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.,* 34(Database issue),D535-9.

Stark,C. et al. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.,* 39, 698-704.

Stelzl,U. et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6), 957-968.

Szklarczyk,D. et al. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research,* 39, D561-D568.

Topol,E.J. et al. (2014). Individualized medicine from Prewomb to Tomb. *Cell,* 157, 241-253

van Driel,M.A. et al. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet,* 14, 535-542.

Vazquez,A. et al. (2003). Global protein function prediction from protein-protein interaction networks. *Nat. Biotech.,* 21(6), 697-700.

Vazquez,A. (2010). "Neuroproteomics". *Taylor and Francis Group, LLC.,* chapter 8.

Wu,X. et al. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol.,* 4, Article 189.

Genetics of Disease - Mendelian and complex disorders. http://medicine.jrank.org/pages/2134/Disease-Genetics-Mendelian-Complex-Disorders.html

국문초록

본 연구는 단백질 상호작용 네트워크의 진화 방법이 생명체가 실제로 존재하고 생명체로서의 역할을 하게 만드는 단백질의 상호작용을 이해하는데 중요하다는 인식으로부터 출발했다. 이러한 단백질 상호작용 네트워크의 진화와 관련하여 많은 연구자들이 진화 모델을 제시해왔고 이러한 모델들 속에서 단백질 상호작용 네트워크의 위상적 성질들이 중요한 역할을 담당해왔다. 단백질 상호작용 네트워크 모델을 제시한 논문 들에서 단백질 상호작용 네트워크의 위상적 성질은 제시된 모델을 검증하는 주요한 방법으로 사용되어왔다. 이러한 단백질 상호작용 네트워크의 진화 모델을 사용해 가상의 초기 단백질 상호작용 네트워크로부터 현재의 단백질 상호작용 네트워크를 얻어내려는 시도가 있어왔고, 또 한편으로는 단백질들의 진화적 연관관계를 통해 단백질의 계통발생 시기를 측정해내려는 시도가 있어왔다. 이러한 연구의 결과로 최근 단백질의 계통발생 시기를 손쉽게 얻을 수 있는 데이터베이스가 구성되었고, 이는 연구자들에게 하여금 자신들이 관심있는 단백질의 계통발생시기를 쉽게 얻을 수 있는 방법을 제공한다는 의미가 있다. 최근 연구에서는 또한 암과 같은 질병이 특정한 계통발생 시기에 속하는 단백질들과 깊이 연관되어 있다는 보고가 있었다. 선행연구들에 비추어 볼 때 단백질 상호작용 네트워크의 위상적 성질은 단백질 상호작용 네트워크의 진화 모델에서 중요한 역할을 담당해왔고 이는 이러한 위상적 성질이 단백질의 성질들 중 진화와 관련된 성질들과 연관되어 있을 것이라는 점을 시사한다.

단백질 상호작용 네트워크의 진화 모델과 단백질의 계통 발생 시기는 단어의 정의상 밀접하게 연관되어 있고, 이러한 사실은 단백질 상호작용 네트워크의 위상적 성질이 단백질의 계통 발생 시기를 측정하는데 사용 될 수 있을 것이라는 단서를 제공한다. 질병과 단백질의 계통발생 시기를 연관시키는 연구들은 또한 질병과 연관된 단백질은 단백질 상호작용 네트워크의 위상적 성질을 사용하여 예측해 보는 연구를 진행하는데 동기를 제공한다.

이러한 연구 동기를 바탕으로 우리는 우선 BioGRID 데이터베이스에서 얻은 인간 단백질 상호작용 네트워크의 각 변에 단백질 상호작용 네트워크상의 삼각형을 기반으로 한 점수를 부여하고 이렇게 부여한 점수를 삼각형 기반 점수로 명명했다. 이 후, 삼각형 기반 점수를 STRING 데이터 베이스에서 제공하는 인간 단백질 상호작용 네트워크의 점수와 비교하는 연구를 진행했다. 또한 삼각형 기반 점수와 단백질의 성질들 중 진화와 관련된 성질들의 연관성을 파악하기 위해 삼각형 기반 점수와 서열 유사성 점수의 비교, 삼각형 기반 점수를 활용한 자기 상호작용 단백질의 예측, 삼각형 기반 점수를 활용하여 단백질 계통발생 시기를 추정하는 연구들을 진행했다. 이러한 연구의 결과로 삼각형 기반 점수가 단백질의 계통발생 시기와 밀접히 연관되어 있다는 결과를 얻었다. 우리는 또한 질병과 연관된 단백질을 삼각형 기반 점수를 활용해 예측하는 연구도 진행했다.

주요어: 단백질 상호작용 네트워크, 질병 연관 단백질, 계통발생 시기
학번: 2015-21610