Statistical Analysis of

Protein Interaction Network Topology

by

Yu-An Dong

Bachelor of Arts
Mathematics, Economics-Statistics, Compute Science
Columbia University, May 1998

Submitted to the Department of Mathematics
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the
Massachusetts Institute of Technology
February 2005

Signature of Author .....
Department of Mathematics
December 16, 2004

Certified by .....
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by .....
Rodolfo Ruben Rosales
Chairman, Applied Mathematics Committee

Accepted by .....
Pavel Etingof
Chairman, Department Committee on Graduate Students

Statistical Analysis of

Protein Interaction Network Topology

by

Yu-An Dong

Submitted to the Department of Mathematics
on December 16, 2004 in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

ABSTRACT

Complex networks arise in diverse areas of natural and social sciences and network topology is a key determinant of such systems. In this work we investigate the protein-protein interaction network of the KSHV herpesvirus, which is the first viral system available, and compare it to a prototypical cellular system.

On the local level, we investigated the relationship between interaction and sequence evolution, functional class, phylogenetic class, and expression profiles. On the global level, we focused on large-scale properties like small-world, scale-free, and attack tolerance. Major differences were discovered between viral and cellular systems, and we were able to pinpoint directions for further investigation, both theoretically and experimentally. New approaches to discover functional associations through interaction patterns were also presented and validated.

To put the KSHV network in the context of host interactions, we were able to predict interactions between KSHV and human proteins and use them to connect the KSHV and human PPI networks. Though simulations, we show that the combined viral-host network is distinct from and superior to equivalent randomly combined networks. Our combined network provides the first-draft of a viral-host system, which is crucial to understanding viral pathogenicity.

In a separate chapter, the results of a project combining experiments and bioinformatics are also presented. We were able to report ~30 new yeast protein-protein interactions and pinpoint the biological significance of some of those interactions. The methodology of yeast two-hybrid itself is also tested and assessed.

Thesis Supervisor: Bonnie Berger, Professor of Applied Mathematics

ACKNOWLEDGEMENTS

I would like to thank Professor Bonnie Berger for introducing me to the field of bioinformatics, mentoring, and support, Alex Coventry and Phil Bradley for their generous help and tutoring while I was getting started, Gopal Ramachandran for discussions and comradeship.

I would also like to thank Dr. Peter Uetz for introducing me to the biologist's way of thinking; his weekly lab meetings were most beneficial in preparing me for a new scientific culture.

I am grateful to Dr. Jürgen Haas for the KSHV project and mentoring. His kindness and support have been crucial to my progress.

Finally, I would like to thank my parents, whose influence on me is omnipresent, and Elena, for being the light. This thesis is dedicated to them.

Chapter 1 Introduction

Complex networks arise in diverse areas of natural and social sciences. For example, the Internet is composed of computers and routers (nodes) connected by physical or telecommunicational links (edges); in a social network, the nodes are individuals and the edges are various social relationships; in a biological cell, proteins physically interact with each other, forming a complex network central to the cell's proper functioning. Due to their huge size and the complexity of their interactions, however, such networks remain poorly understood and their topology remains largely unknown. Are there any organizing principles behind such complex networks? How could they have evolved, that is, what is the interplay between network topology and network dynamics? Can we assess the robustness of such networks and predict their behaviors under perturbation?

Traditionally complex networks have been modeled using the random graph theory of Erdös and Rényi. However, as data on real-world networks accumulate, aided by computerized data acquisition and analysis, it has become increasing clear that the ER model does not fit well with the data – real networks are simply not random! In recent years, emerging topological properties of such complex networks have been discovered and various new models have been proposed [1, 2].

One important complex network is the protein-protein interaction (PPI) network of a cell or a micro-organism. Proteins are the "worker" molecules in a cell, performing diverse roles ranging from structural components to signaling pathways. Proteins rarely act alone, however, rather they interact with each other to accomplish their goals. Thus, in order to understand the functioning of a cell and, more generally, life itself, it is of paramount importance to understand the underlying PPI network.

Unfortunately, PPI network is currently available only for a few selected model organsims [3-5]. Despite all the technological advances, it remains costly and time-consuming to experimentally map protein-protein interactions on a genomic scale. Thus, the available PPI networks not only provide blueprints for their own organisms, but are

invaluable as prototypes, from which general patterns might be discovered and conclusions drawn.

While there exist prototypical PPI networks for eukaryotic cells, some of which have been intensely studied, there has been no systematic result for any of the micro-organisms, in particular viruses. Recently Haas and coworkers have completed the first systematic mapping and analysis of the PPI network for Kaposi Sacoma's Herpesvirus (KSHV), which constitutes the first viral system available [6]. In this thesis we present our work on the bioinformatical analysis of the KSHV network.

In this introductory chapter we give some background on molecular biology, bioinformatics, the herpesviral family, as well as the experimental setup of the KSHV project. Aside from introducing the necessary terminologies, we also give an overview of the larger field so that our work can be put in perspective.

While network topology is one of the most exciting aspects of the KSHV project, there are other aspects that are both important in theory and useful in practice. We systematically address all major aspects of the KSHV network, with a prototypical cellular network from yeast alongside for comparison. The analyses done roughly fall into two categories: local and global. In local analysis, the interactions are treated as a binary dataset, with their network structure (i.e. connection patterns) only in the background. In contrast, large-scale network properties are the main focus of global analysis. Major differences between viral and cellular systems were discovered on both local and global levels, and we were able to pinpoint directions for further research, both theoretically and experimentally, some of which are being actively pursued.

Chapter 2 presents the results from local analysis. We investigate the relationship between interaction and other important characteristics of proteins, including sequence evolution, functional class, phylogenetic class, and expression profile. In addition, we predict viral-viral interactions in other major herpesviruses.

Chapter 3 focuses on the global aspects of the KSHV network. After introducing the necessary background and notations, we investigate the key topological features of the KSHV viral system and compare them to those from a prototypical cellular system. Among the many key differences discovered, we show that

1. Albeit scale-free, the KSHV network has an unusual scaling exponent, which cannot be explained by current dynamic network evolution models and leads to increased attack tolerance.
2. The KSHV network is not small-world, implying many of its interactions are dynamic rather than static.
3. The KSHV network does not exhibit declining degree correlation, which suggests decreased modularity.

In addition to the comparative network analysis of KSHV and yeast, new approaches to discovering functional associations through semi-global interaction patterns are also presented and validated in this chapter.

Since viruses do not act on their own and their pathogenicity is only defined through their interactions with their hosts, we would like to put our KSHV network in the context of host interactions, which is the topic of Chapter 4. We were able to predict interactions between KSHV and human proteins *in silico* and use them to connect the two systems. Furthermore, through simulations we show that our combined viral-host network is distinct from and superior to equivalent, randomly combined networks.

Finally, in Chapter 5 we present the results of a project combining experiments and bioinformatics. We report ~30 new yeast protein-protein interactions and confirm another ~30 previously unreliable ones. The methodology of yeast two-hybrid itself is also tested and assessed, and we show reproducibility is the key to screening out false positives. We develop several independent measures to assess the quality of large-scale PPI datasets. The quality of our own dataset is confirmed, and is compared to that of other genome-wide screens.

1.1 Introduction to Molecular Biology

Cell in a Nutshell

Living organisms consist of cells. Just like the physical world consists of atoms and their interactions, the organic world consists of cells and their interactions. While there are many different types of cells, they all share some common features. A typical eukaryotic cell, in its simplest form, can be visualized as a compartment closed off by cellular membrane ("wall") and filled with fluid (cytoplasm), in which some other smaller compartments reside. The most important of them, the nucleus, is the genetic information storage and control center.

DNA, RNA, Protein and the Central Dogma

DNA, or deoxyribonucleic acid, contains the complete genetic information that defines the structure and function of an organism. DNA consists of two associated polynucleotide strands that wind together in a helical fashion, the famous "double helix".

Each polynucleotide is a linear polymer in which the monomers (deoxynucleotides) are linked together by means of phosphodiester bridges, or bonds. Chemically, each deoxynucleotide consist of a deoxyribose (sugar), a phosphat group ("fuel"), and one of the four types of organic bases. The four bases, Adenine (A), Guanine (G), Cytosine (C), Thymine (T), pair up complementarily on the double-stranded DNA, with As and Gs on one strand paring up with Ts and Cs on the other, respectively. Thus, from an informational point of view, a DNA molecule is a linear sequence over an alphabet of four letters.

To read and execute the genetic instructions contained in DNA, the information is first copied into a messenger molecule, RNA (ribonucleic acid). RNA is similar to DNA in chemical composition; however, unlike the double-stranded DNA, RNA usually consists of a single strand. After being copied ("transcribed") from DNA in the nucleus, RNA

enters the cytoplasm, carrying the information further downstream. On ribosomes the information contained in RNA is read out and translated into the final product, protein.

Proteins are the "worker" molecules in a cell. They catalyze metabolic reactions, transport various "cargos" across the cellular wall and between the organells, receive and relay signals, and form structural components for much of the cell itself.

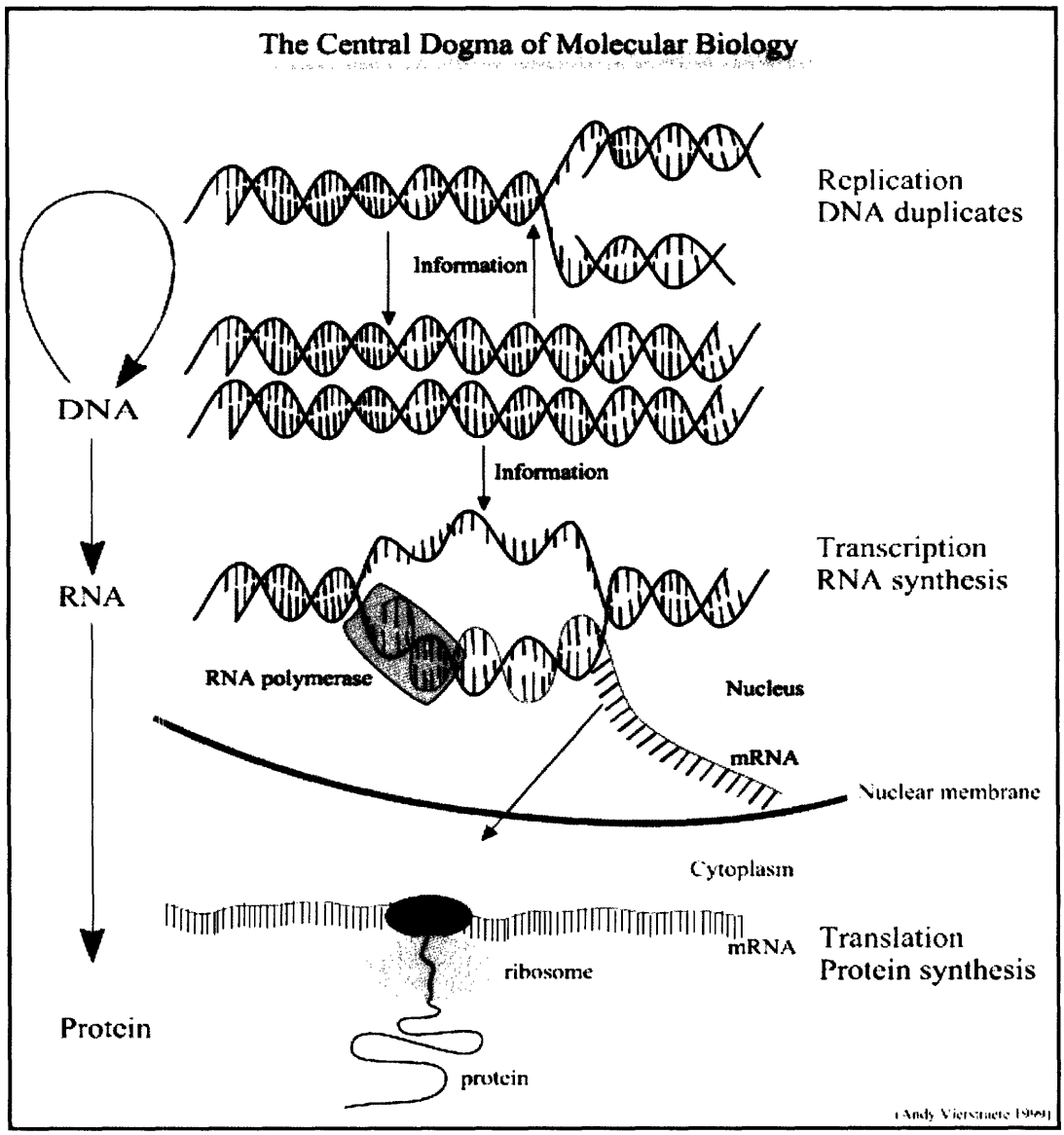
Unlike DNA, which is linear, the structure of a protein can be described on several different levels. Its primary structure is the linear sequence of amino acids connected by peptide bonds, much like DNA being a linear sequence of deoxynucleotides linked together by phosphodiester bonds. There are 20 different amino acids; thus the primary structure of a protein can be viewed as a linear sequence over the alphabet of those 20 letters. Unlike DNA, which stays linear through the two paring strands, local stretches of the protein sequence fold up into well-defined shapes, e.g. alpha-helices and beta-sheets, forming the secondary structure of the protein. Those secondary structural elements then further fold up and pack against one another, forming a compact tertiary structure. It is this 3D structure of proteins that carries out their diverse functions.

This information flow from DNA to RNA to protein is summarized below, the famous "Central Dogma" of modern molecular biology:

1. **Replication**: a double stranded nucleic acid is duplicated to give identical copies. This process perpetuates the genetic information.
2. **Transcription**: a DNA segment that constitutes a gene is read and transcribed into a single stranded sequence of RNA. The RNA moves from the nucleus into the cytoplasm.
3. **Translation**: the RNA sequence is translated into a sequence of amino acids as the protein is formed. During translation, the ribosome reads three bases (a codon) at a time from the RNA and translates them into one amino acid.

In eukaryotic cells, the second step (transcription) is necessary because the genetic material in the nucleus is physically separated from the site of protein synthesis in the cytoplasm in the cell. Therefore, it is not possible to translate DNA directly into protein,

but an intermediary must be made to carry the information from one compartment to another.



1.2 A Primer on Bioinformatics

Computational molecular biology, or bioinformatics, is a vast and expanding field, which lies at the intersection of biology, physics, mathematics, and computer science. Its diverse

areas range from highly theoretical to highly practical. For example, one can prove certain problems are NP-complete, or write custom software for vendor-specific bio-equipment. One can also look at bioinformatics from the perspective of a user versus that of a developer. For example, the popular sequence database searching program BLAST [7] is used by biologists all over the world, experimental and computational alike, to search for homologs of genes or proteins under investigation, while there has been an active research area focusing on BLAST itself, which continually improves the program's performance and expands its applicable domains. In terms of the biological data types it handles, bioinformatics consists of the following three core areas: sequence, structure, and system.

Sequence Alignment and Database Searching

One of the earliest applications of computation to biology is pairwise sequence alignment. Given two related genes (or equivalently their protein products), it is often illuminating to compare them. For example, by comparing genes responsible for genetic diseases from patients and healthy individuals, one can often pinpoint the causal mutations and understand the molecular basis of the diseases. Fortunately, the problem of aligning two sequences arises in diverse fields and has been thoroughly investigated – the optimal match between the two sequences, where some scoring function is maximized, can be found by dynamic programming.

While it is useful to align two related biological sequences, it is even more powerful to align a group of related sequences. Through multiple sequence alignment [8], one might be able to detect major secondary structural elements, differentiate conserved residues, construct phylogenetic trees, or construct a profile to search for new members of the same family.

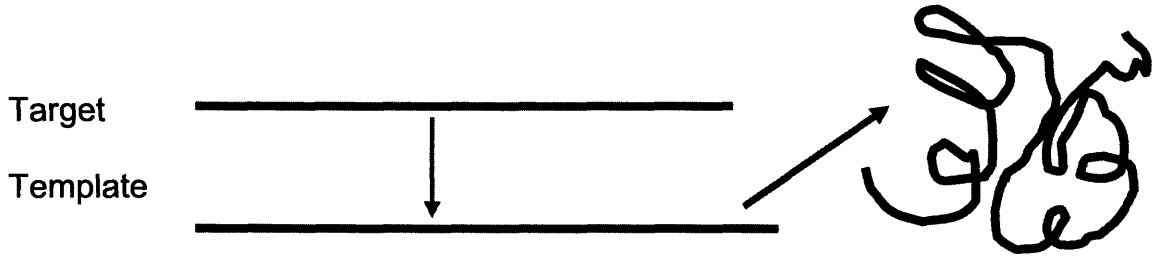
Protein Folding and Structural Prediction

One of the central tenets of modern molecular biology is that sequence determines structure and structure determines function. Thus, one cannot hope to fully understand the function of a protein without knowing its structure. Unfortunately, it is difficult and time-consuming to determine a protein's structure experimentally. Thus, one central problem in molecular biology and one grand challenge in bioinformatics is to predict a protein's 3D structure (target structure) from its linear sequence (target sequence), the so-called protein folding problem.

Despite many years's intense research, the protein folding problem remains open and efforts to predict protein tertiary structure have only met partial success. Depending on whether there are homologous sequences with solved structure and the degree of homology, current structural prediction techniques roughly fall into the following three areas: homology modeling, fold recognition, and *ab initio* prediction.

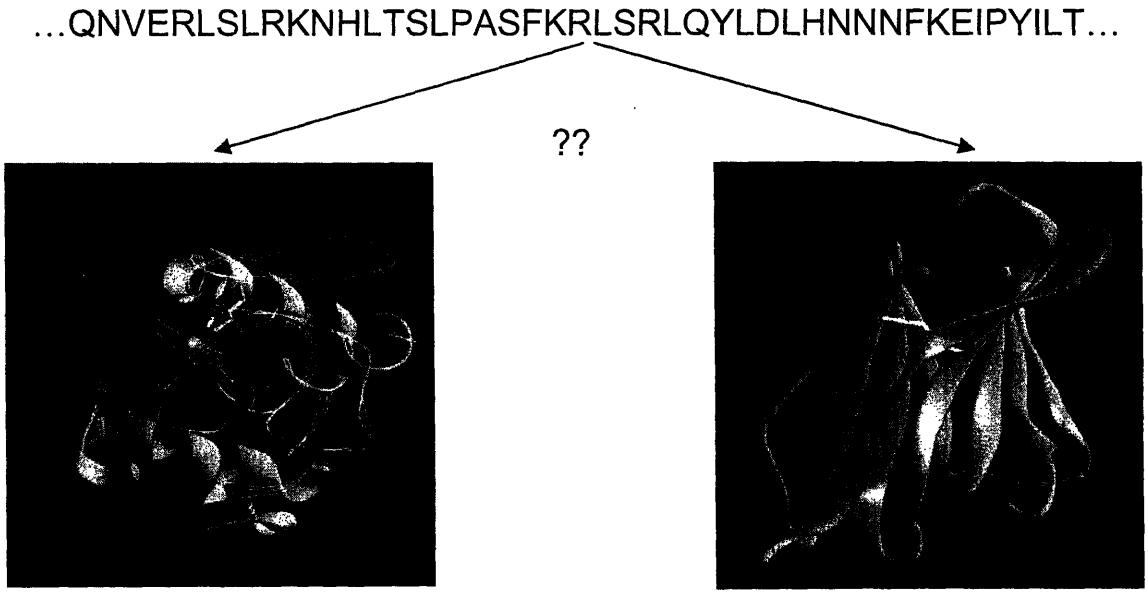
In homology modeling, the target sequence has a homologous sequence with solved structure, and the level of homology between the two sequences is fairly strong (over 40% sequence identity). Then a predicted structure is accomplished in two steps:

1. Align the two sequences
2. Put target sequence onto homologous structure and resolve obvious strains



In fold recognition, or the threading category, the target sequence might share sequence similarity with proteins with known structure, but the homology is not strong enough to make a confident choice among the alternative structures. More generally, threading can

be cast as an inverse folding problem – given a target sequence and a list of alternative structures, predict the structure with the best fit:



Two main approaches in threading are 3D profiles [9] and pairwise contact potentials. While promising in theory, threading certainly has trouble with those multi-domain proteins or those without clearly defined domain structures.

If no homologous structure is available at all, one has to predict the structure from the sequence *ab initio*. The main approach, molecular dynamics, basically recasts the structure prediction problem as an energy minimization problem, with the assumption that the native 3D structure sits at the lowest energy conformation of the given sequence. Unfortunately, the search space of all possible conformations is enormous, even for a protein of moderate length. Indeed, Levinthal has shown that a protein folds up much faster than the time it needs itself to explore all this search space. Thus, there is still something fundamental about the folding process missing in our current understanding – nature does not fold proteins by trying all possible conformations. Aside from philosophical concerns, molecular dynamics also runs into practical obstacles, mainly because we do not have a sufficiently accurate energy function.

In recent years, however, the "logo" method pioneered by Baker and coworkers [10], where small library fragments are assembled into bigger structures, has met with remarkable success.

System Biology and Genome-based Bioinformatics

Since the advent of large-scale genome sequencing projects, a third dimension of bioinformatics, namely system or genome-based biology, has taken on an increasingly important role. Versus traditional, hypothesis-driven biology, system biology represents a paradigm shift. Whole genomes are sequenced, which are the blueprint for cells and organisms. To decipher the complex hierarchy of information, system biology adopts a top-down, discovery-based approach, smoothly integrating experimental and computational aspects. As a first step, coding ORFs can be successfully predicted, either based on gene structure [11] or through comparative genomics. Microarray experiments monitoring mRNA levels in a whole cell, or those associated with a particular condition or process, are carried out. Further down the chain, systematic efforts are being made to map protein-protein interactions on a genomic scale. Other efforts include functional assays like lethality and co-lethality. All this promises to change the face of biology.

As various genome-scale projects progress, there has been an exponential growth of available biological data, which calls for bioinformatics to store and manage them, process and analyze them, integrate and understand them. In theory, the genome of an organism contains all the blueprints to make that organism. The ultimate goal of bioinformatics and biology itself is, in no less measure, to understand this blueprint and life itself.

1.3 A Primer on Herpesvirus

Herpesviruses are wide-spread in mammals and even in some invertebrates. They possess one of the largest viral genomes known. There are three major herpesviral families:

α-herpesvirus, e.g. HSV1

β-herpesvirus, e.g. CMV

γ-herpesvirus, e.g. EBV and KSHV

Four biological properties characterize members of the Herpesviridae family.

- Herpesviruses express a large number of enzymes involved in metabolism of nucleic acid (e.g. thymidine kinase), DNA synthesis (e.g. DNA helicase/primase) and processing of proteins (e.g. protein kinase).
- The synthesis of viral genomes and assembly of capsids occurs in the nucleus.
- Productive viral infection is accompanied by inevitable cell destruction.
- Herpesviruses are able to establish and maintain a latent state in their host and reactivate following cellular stress. Latency involves stable maintanence of the viral genome in the nucleus with limited expression of a small subset of viral genes.

The success of herpesvirus infections depends upon several strategies. The first is the fast efficient way the virion invades the host cell, turning off host protein synthesis and releasing viral DNA into the nucleus, where replication and virion production start immediately. Another strategy that herpesviruses share is the ability to thwart attacks from the host. Tactics include inhibiting splicing of mRNA, blocking presentation of antigenic peptides on the cell surface and blocking the apoptosis (cell death) induced by viral gene expression. A third important strategy shared by herpesviruses is their ability to hide their bare, circularized genome in the nucleus of lymphoma and central nervous system cells and then return to productive infection months, even years later. These latent herpesvirus infections are often benign, but can be devastating to newborns and immuno-suppressed individuals.

1.4 Systematic Mapping of the KSHV Interactome

All KSHV open reading frames (ORFs) were cloned by recombination and the corresponding bait and prey arrays were generated. Since the yeast two-hybrid (Y2H)

system takes place in nucleus and hence is unsuitable for transmembrane proteins, full-length proteins as well as extra- and intracellular domains were cloned separately. To address the known asymmetry between bait and prey in the Y2H system, each pair of proteins is tested in both directions for interaction. In total, over 12000 interactions, corresponding to all possible bait-prey combinations, were tested as a matrix. Among them, 123 unique interactions were identified.

Since Y2H is known to generate a large number of false positives, all positive Y2H interactions were retested under both β-galactosidase assay (Gal) and co-immunoprecipitation (CoIP). Approximately 50% of the Y2H interactions were confirmed by CoIP.

Chapter 2 Local Analysis

2.1 Prediction of viral-viral interactions in other herpesviruses

2.1.1 Introduction

To date there has been no comprehensive, large-scale study on viral protein interactions, be it viral-viral or viral-host. Currently known interactions have been generated by small-scale, individual experiments; as such, the coverage is both limited and biased – there are not many reported interactions, and the great majority of them focus on viral-host and on well-known proteins or processes. For example, we could only find 3 reported viral-viral interactions in KSHV itself after scanning more than 1000 PubMed abstracts on this topic. Since we have obtained 123 KSHV viral-viral interactions, roughly half of which have been confirmed by high-confidence CoIP experiments, we would like to extend this knowledge to other major herpesviruses by predicting their viral-viral interactions in silico, hence generate first-draft viral-viral interaction networks for them – biologists could then experimentally verify the predicted interactions with priority and already start to make/validate hypotheses on those predicted networks.

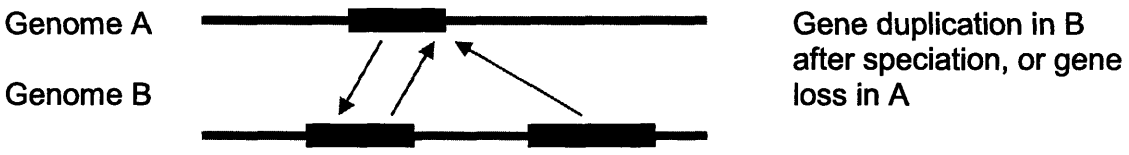
2.1.2 Identification of KSHV orthologs in other herpesviruses

Biologically, orthologs can be defined as genes in different organisms that are direct evolutionary counterparts of each other, which arise through speciation. Orthologs are believed to perform the same function and have the same specificity, that is, the same interaction partners, if one thinks of them in terms of their protein products. On the other hand, paralogs are related genes in a single organism that arise through gene duplication. Being a duplicate of the original gene, a paralog would be under less evolutionary pressure to maintain the same specificity, that is, it has more latitude to evolve and acquire new functions. In terms of protein products, this would mean loss of some old interaction partners and gain of new ones. Thus, if we could successfully identify KSHV

orthologs in other herpesviruses and distinguish them from paralogs, then, under the assumption that orthologs have the same functional specificity and interaction partners, we would be able to predict orthologous interactions in other herpesviruses. For example, if k1 and k2 interact in KSHV and they have orthologs h1 and h2, respectively, in another herpesvirus H, then we predict h1 and h2 interact in H.

While the biological definition of orthologs is the correct one, it is not particularly easy to apply in practice – to identify an ortholog, one would need such detailed knowledge of its biological function and evolutionary history, which is feasible for only a handful of most well-studied genes.

Operationally, one can define orthologs as reciprocal best BLAST hits in two proteomes. For example, if protein a in organism A has protein b in organism B as top hit (that is, when one runs BLAST using a as query and B as database, b turns up as the one with the best e-value) and vice versa, then we consider a and b as a pair of orthologs.



Complete genomes of KSHV(K), HSV1(A), CMV(B), and EBV(C) were downloaded from GenBank at the NCBI website. For each genome file, all unique protein sequences corresponding to CDS entries were extracted and compiled into a BLAST database. All pairs of databases were searched against each other through stand-alone BLAST program, with cutoff e-value set at 0.1, and reciprocal best hits (i.e. orthologs) were extracted. Python scripts were used to automate the above process.

2.1.3 Results

ORF K1	***	***	***	HI
ORF 4	***	UL32	BLLF1b	HI
ORF 6	UL29	UL57	BALF2	DN
ORF 7	UL28	UL56	BALF3	VS
ORF 8	UL27	UL55	BALF4	VS
ORF 9	UL30	UL54	BALF5	DN
ORF 10	***	***	***	UN
ORF 11	***	***	***	UN
ORF K2	***	***	***	HI
ORF 2	***	***	***	DN
ORF K3	***	***	***	HI
ORF 70	***	***	***	DN
ORF K4	***	***	***	HI
ORF K4.1	***	***	***	HI
ORF K4.2	***	***	***	UN
ORF K5	***	***	***	HI
ORF K6	***	***	***	HI
ORF K7	***	***	***	HI
ORF 16	***	***	***	HI
ORF 17	UL26	UL80	BVRF2	VS
ORF 18	***	UL79	Predicted	UN
ORF 19	UL25	UL77	BVRF1	VS
ORF 20	UL24	UL76	BXRF1	UN
ORF 21	UL23	***	BXLF1	DN
ORF 22	UL22	UL75	BXLF2	VS
ORF 23	***	UL117	BTRF1	UN
ORF 24	***	UL87	BcRF1	UN
ORF 25	UL19	UL86	BcLF1	VS
ORF 26	UL18	UL85	BDLF1	VS
ORF 27	***	***	BDLF2	UN
ORF 28	***	***	BDLF3	UN
ORF 29b	UL15	UL89	BDRF1	VS
ORF 30	***	***	***	UN
ORF 31	***	UL92	BDLF4	UN
ORF 32	UL17	UL93	BGLF1	VS
ORF 33	UL16	UL94	BGLF2	VS
ORF 29a	***	***	BGRF1	VS
ORF 34	***	UL95	BGLF3	VS
ORF 35	***	***	***	UN
ORF 36	UL13	UL97	BGLF4	HI
ORF 37	UL12	UL98	BGLF5	DN
ORF 38	***	***	***	VS
ORF 39	UL10	UL100	BBRF3	VS
ORF 40	***	***	BBLF2	DN
ORF 41	***	***	***	DN
ORF 42	UL7	UL103	BBRF2	DN
ORF 43	UL6	UL104	BBRF1	VS
ORF 44	UL5	UL105	BBLF4	DN
ORF 45	***	***	BKRF4	GR
ORF 46	UL2	UL114	BKRF3	DN
ORF 47	***	***	BKRF2	VS
ORF 48	***	***	***	VS
ORF 49	***	***	BRRF1	UN

ORF 50	***	***	BRLF1	GR
ORF K8	***	***	***	GR
ORF K8.1	***	***	***	VS
ORF 52	***	***	BLRF2	UN
ORF 53	***	UL73	BLRF1	VS
ORF 54	UL50	UL72	BLLF3	DN
ORF 55	***	***	BSRF1	UN
ORF 56	UL52	UL70	BSLF1	DN
ORF 57	UL54	UL69	BMLF1	GR
ORF K9	***	***	***	HI
ORF K10	***	***	***	HI
ORF K10.5	***	***	***	HI
ORF K11	***	***	***	HI
ORF 58	***	***	BMRF2	UN
ORF 59	***	***	BMRF1	DN
ORF 60	UL40	***	BaRF1	DN
ORF 61	UL39	UL45	BORF2	DN
ORF 62	***	***	BORF1	VS
ORF 63	***	***	BOLF1	VS
ORF 64	UL36	UL48	BPLF1	VS
ORF 65	***	***	BFRF3	VS
ORF 66	***	UL49	BFRF2	UN
ORF 67	UL34	UL50	BFRF1	VS
ORF 67.5	UL33	UL51	BFRF4	VS
ORF 68	UL32	UL52	BFLF1	VS
ORF 69	UL31	UL53	BFLF2	VS
ORF K12	***	***	***	HI
ORF K13	***	***	***	HI
ORF 72	***	***	***	HI
ORF 73	***	***	***	HI
ORF K14	***	***	BARF1	HI
ORF 74	***	US28	***	HI
ORF 75	***	***	BNRF1	VS
ORF K15	***	***	***	HI

19 Predicted Interactions in HSV1

UL30	UL33
UL15	UL33
UL13	UL39
UL52	UL13
UL54	UL54
UL54	UL39
UL54	UL32
UL40	UL52
UL40	UL40
UL40	UL39
UL40	UL33
UL40	UL32
UL31	UL33
UL30	UL10
UL30	UL32

UL15	UL50
UL15	UL32
UL13	UL50
UL39	UL39

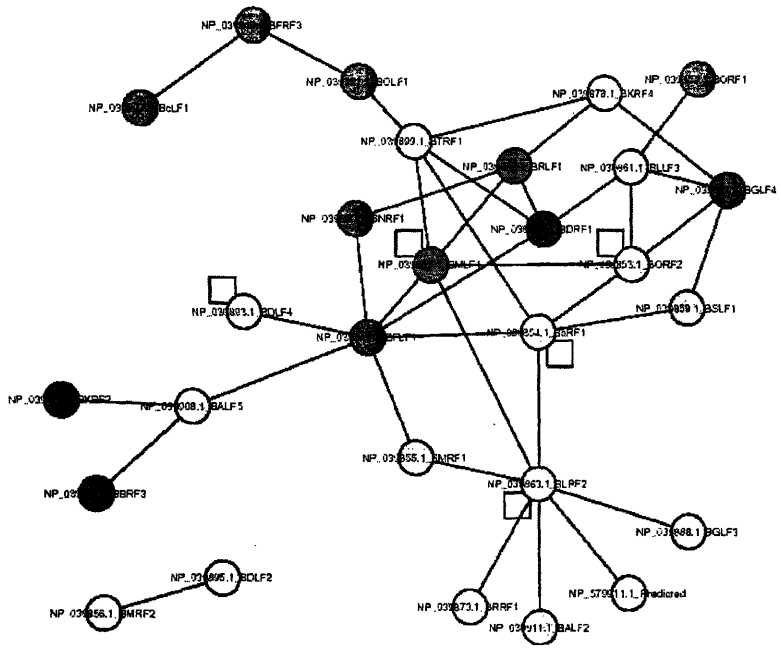
22 Predicted Interactions in CMV

UL54	UL51
UL117	UL51
UL89	UL117
UL89	UL51
UL92	UL92
UL92	UL51
UL92	UL52
UL97	UL45
UL70	UL97
UL69	UL117
UL69	UL69
UL69	UL45
UL69	UL52
UL53	UL51
UL54	UL100
UL54	UL52
UL89	UL72
UL89	UL52
UL89	US28
UL95	UL51
UL97	UL72
UL45	UL45

43 Predicted Interactions in EBV

BALF5 BFRF4
BTRF1 BFRF4
BcLF1 BFRF3
BDRF1 BTRF1
BDRF1 BDLF3
BDRF1 BFRF4
BDLF4 BDLF4
BDLF4 BFRF4
BDLF4 BFLF1
BGLF4 BORF2
BLRF2 BLRF2
BLLF3 BORF1
BSLF1 BGLF4
BMLF1 BTRF1
BMLF1 BMLF1
BMLF1 BORF2

BMLF1 BFLF1
BMRF1 BFLF1
BaRF1 BSLF1
BaRF1 BaRF1
BaRF1 BORF2
BaRF1 BFRF4
BaRF1 BFLF1
BOLF1 BTRF1
BOLF1 BFRF3
BOLF1 BFRF4
BFLF2 BFRF4
BNRF1 BRLF1
BNRF1 BFRF4
BNRF1 BFLF1
BALF2 BLRF2
BALF5 BBRF3
BALF5 BKRF2
BALF5 BFLF1
BTRF1 BDLF3
BTRF1 BKRF4
BDLF3 BDLF3
BDLF3 BFRF4
BDRF1 BRLF1
BDRF1 BLLF3
BDRF1 BFLF1
BGLF3 BLRF2
BGLF3 BFRF4
BGLF4 BKRF4
BGLF4 BLLF3
BKRF4 BRLF1
BRRF1 BLRF2
BMLF1 BRLF1
BMLF1 BLRF2
BMRF2 BDLF2
BMRF1 BLRF2
BMRF1 BFRF4
BaRF1 BTRF1
BaRF1 BLRF2
BORF2 BORF2
BFLF2 BLRF2



2.2 *Sequence Evolution*

2.2.1 Motivation

Sequence alignment programs constitute a major contribution of bioinformatics to biology and are routinely used by biologists all over the world, often as the first step in analyzing the gene or protein of interest. In the case of proteins, when two or more

related sequences are properly aligned, conserved residues and regions can be readily identified, which often correspond to elements important for function, structure, or folding kinetics. Moreover, the "texture" of the alignment, i.e. the alternating spacing between conserved and non-conserved regions often shed light on possible 3D structure and domain composition. The overall percentage identity in a pairwise alignment is used to measure how closely related the two underlying sequences are, in the sense of molecular evolution.

All life forms on earth come from a single ancestor. While the myriad of living organisms are at once colorful and confusing, they all consist of cells. A cell is the smallest unit that can be considered "alive", i.e. independently capable of growth and reproduction (unlike a virus, which must rely on host organisms for replication, hence infection). Cells, of the same type or heterogeneous, cooperate and interact to form organs and the organism itself, much like how atoms interact to form the physical world. Thus to understand life, one must first understand cell.

When a cell replicates, it must copy its own genome and pass it along to the daughter cell. This, however, is an inherently noisy process and stochastic errors abound. Most copying errors in a gene, called mutations, are fatal – they either make the daughter cell immediately inviable or make it less fit, so that this line of cells will be less able to compete with normal, healthier cells and their progenies and will be gradually wiped out. Occasionally, however, a mutation can be neutral or even beneficial to the daughter cell and hence establish itself in the population. Living in different environments, cells are free to accumulate beneficial mutations unique to their respective environments. When enough mutations exist between two cell lines, they can be considered different species.

Different genes (respectively proteins) accumulate mutations at different rate. Some genes are crucial to cell viability and hardly any deleterious mutations will be tolerated, while others are under less evolutionary pressure and have more latitude to accumulate mutations. For example, the histone proteins play a critical role in the package of DNA within the nucleus. All four core histones (H2A, H2B, H3 and H4) are highly conserved,

with H4 having more than 95% sequence identity across all known H4 sequences, from yeast to human! At the other end of the spectrum, conservation can drop to below 20% for distantly related species.

For a single gene (respectively protein), different regions also accumulate mutations at different rate. For example, residues on protein interface tend to be more conserved than the rest of residues in a protein sequence, so do ligand-binding and active sites. Since most proteins carry out their cellular functions through protein-protein interactions, one would expect PPI to put a constraint on protein sequence evolution. Indeed, it has been shown for the yeast cell that interacting proteins are more conserved (i.e. have higher sequence identity) than those with no known interaction partners [12]. Furthermore, there is a positive correlation between the number of interaction partners a protein has and its degree of conservation [13].

Since KSHV is the first viral system with enough PPIs available, we would like to investigate the relationship between PPI and protein sequence conservation, and compare our results to those from a cellular system.

2.2.2 Methods and Results

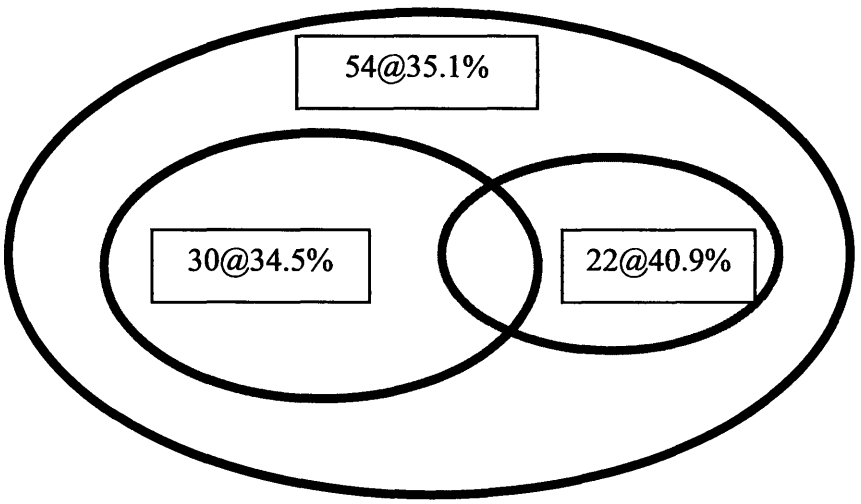
Using reciprocal best BLAST hits, we have identified putative KSHV orthologs in HSV1, CMV, and EBV. Given the large evolutionary distance separating the herpesviral families, local alignment is much more appropriate than global alignment. Thus we have taken the BLAST % identity as the measure for sequence conservation between KSHV ORFs and their orthologs, in hope that this most conserved region contains the key functional domain or protein-protein interaction interface. In general, we found that herpesviruses are fairly divergent – many ORFs do not have orthologs and the homology (% identity) is generally low even if they do.

Among the 83 KSHV ORFs, 54 have orthologs in EBV, with an average sequence identity at 35.1%. Among those 54 ORFs, 30 have interactions in our screen, with an

average sequence identity at 34.5%. Thus interacting proteins are not more conserved in KSHV, in contrast to cells. We hypothesize that this discrepancy is largely due to "hidden" host interactions, where some KSHV proteins target many host proteins but do not have viral interaction partners.

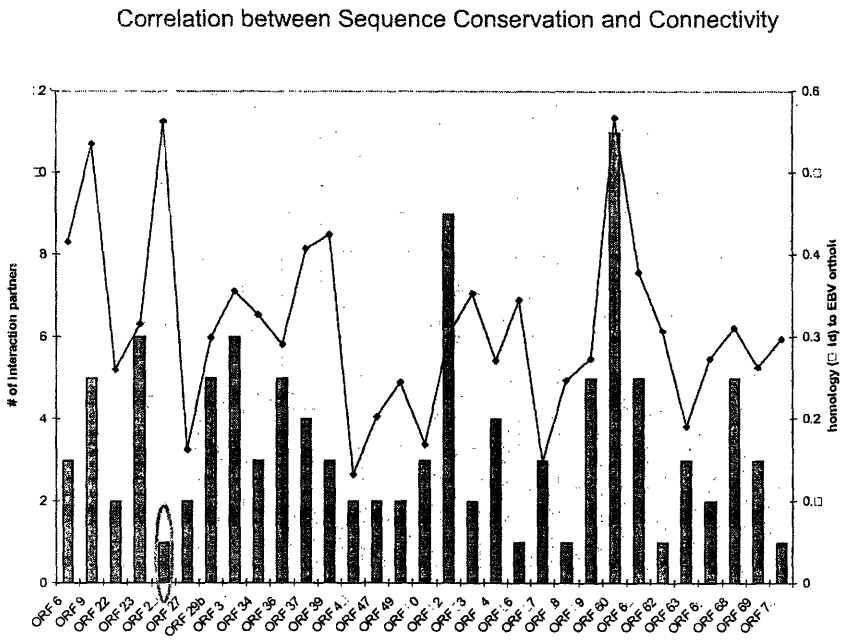
On the other hand, among those 54 KSHV ORFs, 22 have orthologs all across the three major herpesviral families, that is, of phylogenetic class 111. This core set of proteins are, however, more conserved, with an average sequence identity at 40.9% ($p = 0.027$ under t-test).

The following Venn diagram summarizes this relationship:



However, the constraint put on sequence evolution by viral PPI should still exhibit itself, once we factor out the hidden effect of viral-host interactions. Among the 54 KSHV ORFs with orthologs in EBV, 30 have viral-viral interaction partners in our screen. Now we investigate the relationship between sequence conservation and the number of interaction partners for those 30 KSHV ORFs. In contrast to the previous analysis, where we compared KSHV ORFs with viral-viral interactions to those without, now we compare the former set of ORFs among themselves. The idea behind this is that proteins

with and without viral interaction partners represent two distinct classes, where hidden effect like host interactions could play a major role; on the other hand, within the class of proteins with viral interactions, the effect of host interactions, even if still present, would apply in roughly equal measure to all members and cancel each other out, provide there is no systematic bias.

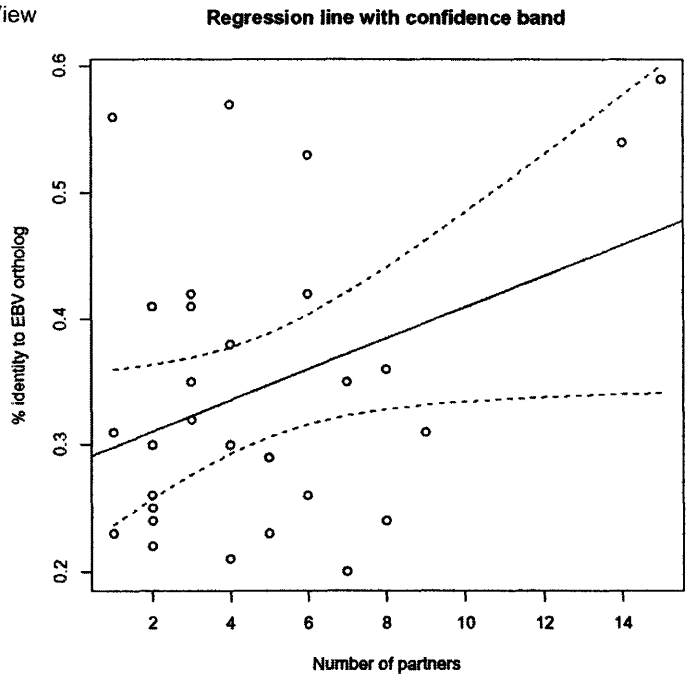


For the 30 KSHV ORFs with orthologs in EBV, we plotted the sequence identity (lines) together with the number of interaction partners (bars). Indeed we observe a significant positive correlation between sequence conservation and connectivity, with $r = 0.368$, $p = 0.046$. Thus, hubs are indeed more conserved.

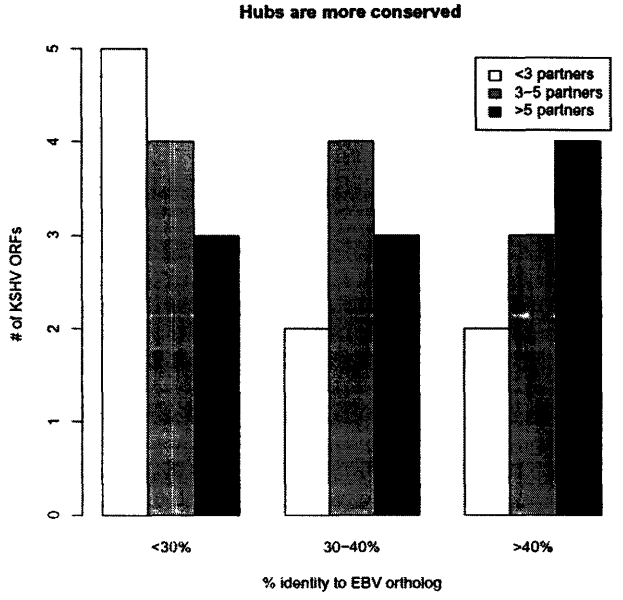
Aside from proving this important correlation, the analysis also pinpoints interesting exceptions where there is high homology but low number of interactions and vice versa. A good example would be ORF 25 (3rd highest homology but only 1 partner); ORF 57 would be another example (lowest homology but with 7 partners). Reassuringly, ORF 25 is the major capsid protein, a key structural protein in virus shell assembly. Thus it has

only a small number of interaction partners because its interaction is rather specific, and it is highly conserved because it needs to maintain a precise 3D structure for assembly. Here are two other views of this correlation:

Another View



Yet Another View



- Observe trends
- Identify exceptions

2.3 *Interactions among Functional Classes*

2.3.1 Motivation

Proteins are the operational molecules in a cell and play diverse roles from enzymes to structural components. However, proteins rarely act alone – rather they act in close coordination, through protein-protein interactions, to accomplish their goals. For example, most cellular machineries are protein complexes, several proteins held tightly together by stable PPI, while a series of transient PPIs, where proteins briefly associate and then dissociate, are responsible for signalling cascades. Thus it is reasonable to assume that interacting proteins participate in related biological processes and share similar biological functions (i.e. cellular roles), though their exact biochemical functions may differ. Indeed, Schwikowski et al [3] has shown that, for a large, high-confidence set of yeast PPIs, interacting proteins are more likely to share a function than random pairs of proteins.

Conversely, one can use this observation to assess the quality of PPI datasets. It has been shown that there is a considerable difference between PPI datasets compiled from individual publications and those obtained from genome-scale experiments, and argued that those genome-wide datasets contain a huge number of false positives [14].

On the predictive side, one can assign tentative functions to a protein of unknown function based on those of its interaction partners. This "guilt-by-association" approach has proven successful.

Finally, the crosstalks between and within functional classes may provide biological insights.

2.3.2 Methods and Results

Given the importance of understanding interactions among functional classes, we would like to investigate this topic for KSHV, for the first time a non-cellular system.

Unfortunately, knowledge of KSHV protein functions is still rudimentary and scattered in literature – there has been no comprehensive functional classification to date, with most proteins assigned "unknown" functions. To circumvent this problem, we looked at the GenBank annotations for each KSHV ORF and those for its orthologs in other herpesviruses, if available, and were able to assign a function to most of them with reasonable confidence. After some further adjustments based on complementary literature, the KSHV ORFs are partitioned into five broad functional classes:

'DN' = DNA replication, nucleotide metabolism
'GR' = gene regulation
'HI' = host interaction
'UN' = unknown
'VS' = virion structure

There are 123 unique interactions among 50 KSHV proteins. Since self-interactors always share the same function and bias the result, they are removed from the dataset and we are left with 115 interactions among 50 proteins. The distribution among functional classes is summarized in the following table:

	DN(10)	GR(4)	HI(13)	UN(8)	VS(15)
DN	5	3	22	6	12
GR		4	2	3	3
HI			6	8	19
UN				9	10
VS					11

To estimate their statistical significance, we first introduce some notations. Suppose there are E interactions (edges) among N proteins (nodes), which fall into C functional classes. Let $n_1, n_2, ..., n_C$ be the counts of nodes in the functional classes, where $\sum n_i = N$. Let p_{ij} be the probability of interaction between functional classes i and j. Then we have

$p_{ij} = \begin{cases} 2 p_i p_j, i \neq j \\ p_i p_j, i = j \end{cases}$, where $p_i = \frac{n_i}{N}$ is the probability of picking a node from functional class i. Let q_{ij} be the observed (relative) frequency of interaction between functional classes i and j, which is simply the raw count of such edges e_{ij} divided by E.

Now we define the odds ratio to be $o_{ij} = q_{ij} / p_{ij}$, that is, observed frequency over background frequency, and use it to measure the over- and under-representation of interactions among functional classes.

As a complementary measure, we also directly compute a p-value for the observed count of edges e_{ij} between functional classes i and j. Let X_{ij} be the number of such edges, then X_{ij} is a binomial random variable, with E as the number of trials and p_{ij} as success probability. The p-value is then computed as $P_{ij} = \lfloor X_{ij} \leq e_{ij} \rfloor$, that is, the probability of observing at most e_{ij} edges by chance.

The results are summarized in the following table. For each entry, the number before '*' is the odds ratio, the number after being the p-value.

	DN	GR	HI	UN	VS
DN	0.652*0.320	0.815*0.496	1.839*0.998	0.815*0.391	0.869*0.366
GR		4.076*0.993	0.418*0.138	1.019*0.659	0.543*0.192
HI			0.514*0.105	0.836*0.375	1.059*0.664
UN				2.038*0.971	0.905*0.448
VS					1.062*0.660

2.3.3 Discussions

In general one would expect that proteins from the same functional class are more likely to have interactions. In our case, we have:

GR-GR significantly more interactions
UN-UN significantly more interactions
DN-HI significantly more interactions

Some observations:

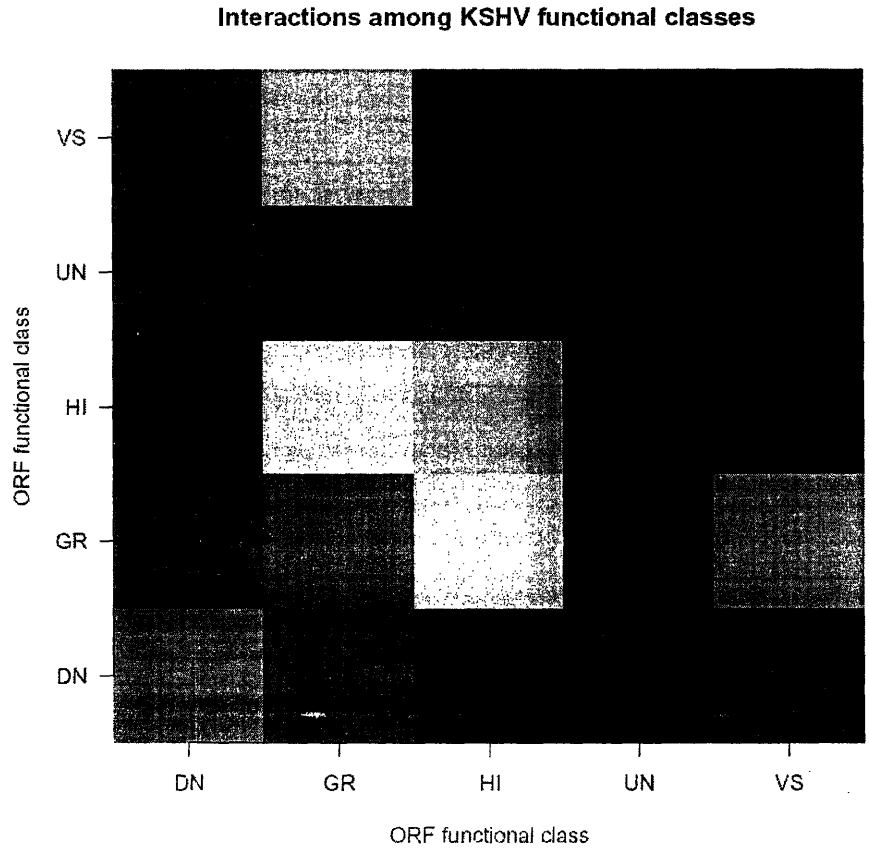
1) DN-DN, VS-VS have roughly the same level of interactions as background, while HI-HI has less. This somewhat makes sense, since DN proteins interact with DNA, VS proteins have specific interactions (e.g. in shell assembly), while HI proteins interact with host.

2) UN-UN has significantly more interactions. This is actually very interesting -- if this class were fairly mixed, i.e. with many ORFs from the 4 true functional classes, UN-UN would look like the background. Thus this implies either ORFs in UN are mostly of the same (yet unknown) function, and/or they form their own complexes/processes. In other words, they are mostly not "the missing parts" of known complexes/processes, but are from entirely new, yet unknown complexes/processes.

3) DN-HI has significantly more interactions -- just want to mention there are 6 ORFs in HI could have been assigned as GR.

Due to the unusual interaction patterns among functional classes, we come to address the fundamental question – are interacting proteins in KSHV, a viral system, more likely to share the same function, as is the case for cellular systems? Of the 115 interacting KSHV proteins, 23.5% fall into the same functional class – this compares to 21.4% for random pairs of KSHV proteins that have interactions. The odds ratio is only 1.098 (i.e. only

roughly a 1.1 fold enrichment), which might not be significant. To construct a valid null model for comparison, it is not enough to only consider the background frequencies of the underlying functional classes – one must also take into account network topology, the connectivity patterns of the interacting proteins as a whole. To satisfy both constraints at once, we fix the topology of the real KSHV interaction network, but randomly permute node labels, i.e. function assignment of proteins. Since the space of such randomly permuted networks is huge, we sample 1000 of them and derive an empirical p-value for the real network. Indeed, the odds ratio is insignificant compared to the random ensemble, with empirical $p = 0.334$. Thus, unlike cellular systems, in KSHV interacting proteins are not more likely to share the same function than random pairs of proteins.



2.4 Phylogenetic Classes

2.4.1 Methods and Results

Just like proteins involved in the same pathway or complex are more likely to interact, they also tend to co-evolve during evolution – all such functionally related proteins tend to be either preserved or eliminated all together in a new species. Conversely, proteins with similar presence/absence patterns in other genomes, or phylogenetic profiles, tend to be functionally linked, in particular more likely to interact [15]. Since the relationship between interaction and functional classes for cellular systems cannot be transferred to KSHV, we would also like to investigate the relationship between interaction and phylogenetic classes for KSHV.

For each KSHV ORF, we encode its phylogenetic profile in a 3-digit binary string, where 0/1 denotes the absence/presence of an ortholog in HSV1, CMV, and EBV, respectively. Thus the phylogenetic profiles can be read off directly from the table of KSHV orthologs. Here we reproduce the top part of that table for illustration:

KSHV	HSV1	CMV	EBV	Phylogenetic Class
ORF K1	***	***	***	000
ORF 4	***	UL32	BLLF1b	011
ORF 6	UL29	UL57	BALF2	111
ORF 7	UL28	UL56	BALF3	111

Recall that there are three major herpesviral families, $\alpha, \beta, and\ \gamma$, with HSV1, CMV, and EBV as representatives, respectively, and that KSHV itself belongs to the γ family. Thus the phylogenetic profiles as we defined have intuitive biological interpretations. For example, proteins of phylogenetic class 000 are KSHV-specific and presumably define its pathogenicity; those of class 001 are unique to the γ family; while those of class 111 are a core set of proteins common to all herpesviruses – presumably they are the most ancient proteins and perform the most fundamental tasks in herpesviruses.

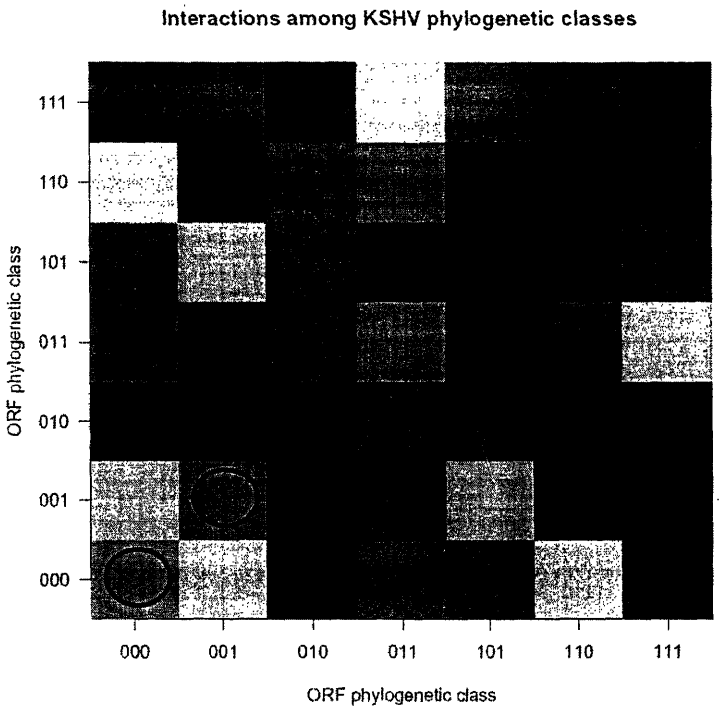
In parallel to the analysis on functional classes, now we similarly compute the over- and underrepresentation of interactions among phylogenetic classes. The table below summarizes the result (the entries are of the form logodds*p-value):

	000	001	011	111
000	0.492*0.021	0.689*0.065	1.494*0.927	1.067*0.690
001		0.962*0.547	0.781*0.416	0.893*0.369
011			1.270*0.813	1.088*0.686
111				1.451*0.941

Again one would expect proteins from the same class to have more interactions (people actually predict interactions by such phylogenetic profiles). But this is NOT true for our data! We have:

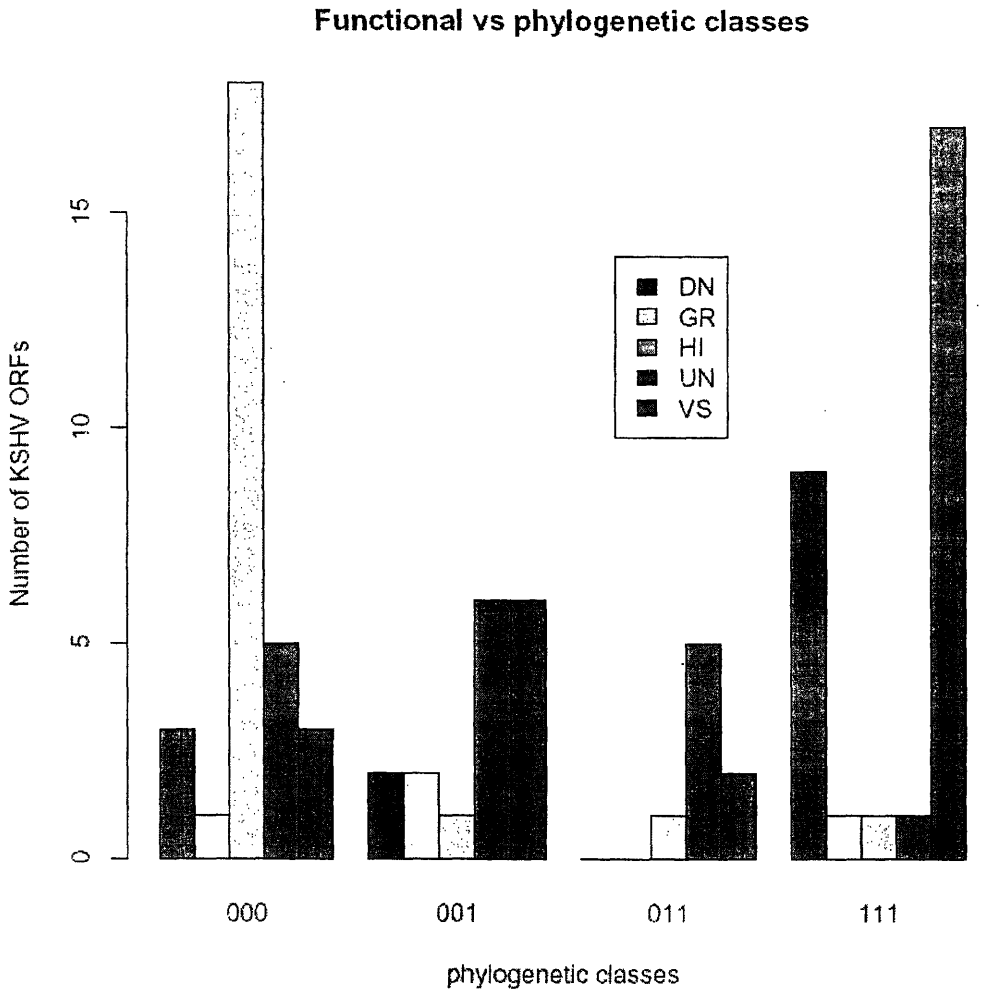
000-000	significant under
000-001	borderline under
000-011	borderline over
111-111	borderline over

Thus KSHV-specific proteins (000) tend to avoid each other, while those proteins conserved in all A, B, C (111) tend to interact with each other.



2.4.2 Further Analysis

After analyses done on both functional and phylogenetic classes, now we look at their intersections. Indeed, quite a few interesting insights can be obtained from the following figure. For example, the class 000 of KSHV-specific proteins are dominated by those involved in host interaction, while the class 111 of proteins conserved all across the three major herpesviral families mostly consists of structural proteins and those involved in DNA replication, both perform basic, fundamental service to the micro-organisms.

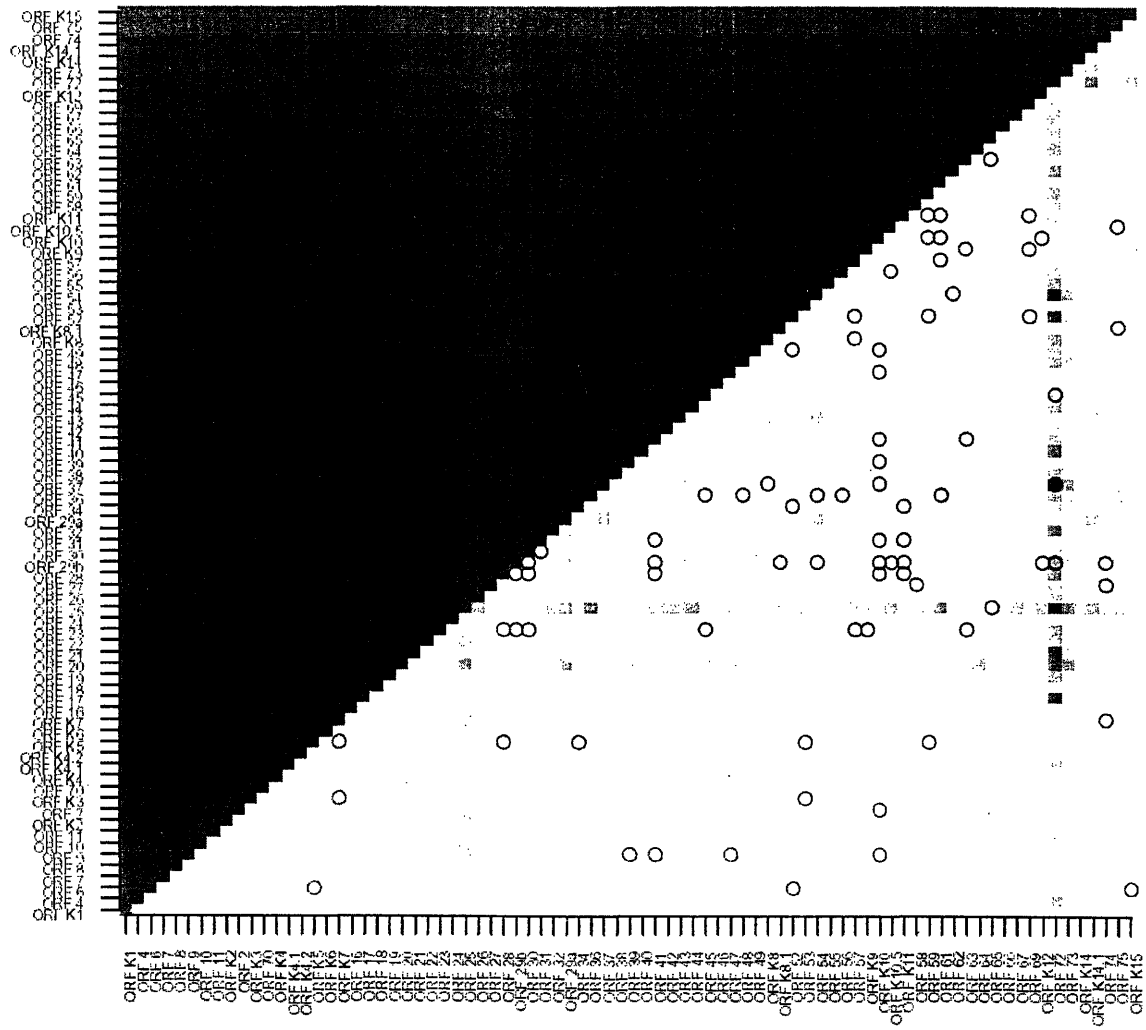


2.5 *Expression Correlation and Interaction*

2.5.1 Y2H versus Random

If two proteins interact, then their expression profiles may correlate. Indeed, this property has been used to assess the quality of genome-scale protein-protein interaction data [14]. In this section, we use expression correlation to validate our own interaction data and to assess the experimental procedures we used.

Expression Correlation of KSHV



circles represent interactions

All pairwise expression profile correlations for 81 KSHV ORFs were plotted in a matrix. The ORFs on either axis are in their order along the KSHV genome. The square at position (i, j) corresponds to the correlation of the expression profiles of ORF i and ORF j, while the intensity of the heat map (going from red to white) corresponds to the magnitude of the correlation (from low to high). This way one can easily spot outliers like ORF 72, which has a very different expression profile from most other ORFs, and

regions sharing similar profiles (those white regions), e.g. the cluster from ORF 7 to ORF 16 shows profile similarity both within itself and with several late clusters.

Our own interactions are plotted as circles in the matrix. This way one can spot clusters, where several ORFs interacting with one another, or one ORF interacting with several other ORFs that are consecutive along the genome.

While intuitive and useful, the matrix does not yield obvious conclusions on the relationship between expression and interaction. Now we do so numerically by computing the correlation of all pairwise expression profiles, with the average being 0.804, which is the background or random correlation. Now we compute the expression correlation of those interacting ORFs, obtaining an average of 0.839 (the difference is significant at p=0.0004). Thus, despite limited sample size and expression profiling condition, interacting proteins in KSHV are indeed expressed at similar time points.

2.5.2 CoIP versus Gal

Since our experiment is the first study where yeast two-hybrid data is comprehensively retested using co-immunoprecipitation and β-galactosidase assays, we would like to assess their relative effectiveness. Again, we use average expression correlation (AEC) as an independent measure to assess the three experimental methods.

Of the 123 Y2H interactions we have, we require:

1) Both ORFs in the pair have expression data.
2) Self-interactions are excluded (since the expressions would correlate perfectly).
3) ORF 72 is excluded, since it is a huge outlier in terms of expression.

After the filtering step, we are left with 77 Y2H interactions, with average expression correlation AEC=0.839. As control, under the same three constraints, the background AEC=0.804.

Now we partition those 77 Y2H interactions into those confirmed by CoIP (CoIP+) and the rest (CoIP-). We have:

Dataset	Size	AEC
CoIP+	36	0.858
CoIP-	41	0.822

The difference between CoIP+ and CoIP- has p=0.05, while the difference between CoIP- and background has p=0.18. Thus CoIP has indeed significantly enriched the Y2H data for true interactions, but we cannot entirely discount those interactions not confirmed by CoIP – there are likely still a significant fraction of true interactions among them.

Similarly, we can also partition the same 77 Y2H pairs according to their Gal level (going from 0, the lowest level, to 3, the highest):

Dataset	Size	AEC
G0	46	0.834
G1	28	0.851
G2	3	0.803

There are no interactions in the G3 class. The difference between G0 and G1 has p=0.36 and thus is insignificant.

Taken together, it seems that CoIP is better than Gal at picking out "true" interactions, if expression correlation we used is a reliable measure. CoIP nicely separates the 77 Y2H interactions into a high-confidence set and the rest, while Gal barely does so. (G0 has about the same AEC as Y2H, while G2 is actually much worse!) Of course, larger datasets and more experiments are needed to conclusively confirm our observations.

Now we look at the intersections of CoIP and Gal to gain further insight:

Size/AEC	CoIP+	CoIP-
G0	20/0.849	26/0.821
G1	15/0.871	13/0.828
G2	1/0.818	2/0.796

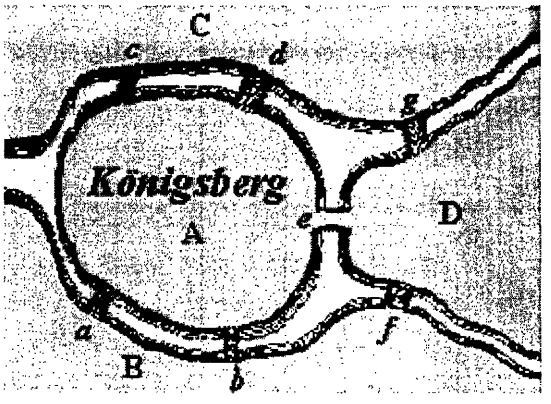
Note Gal further enriches the results of CoIP+ (first column), with CoIP+&G1 having AEC=0.871, but is not effective on CoIP- (second column); on the other hand, CoIP always enriches the results of Gal (all three rows). This would have important implications when we combine experiments to obtain high-confidence interactions.

Chapter 3 Global Analysis

3.1 Background

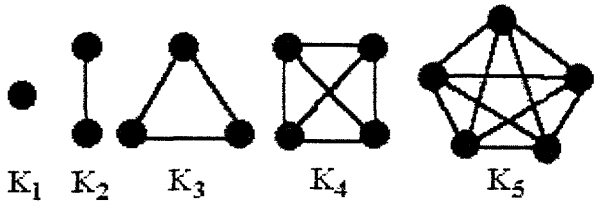
3.1.1 From Regular Graphs to Complex Networks

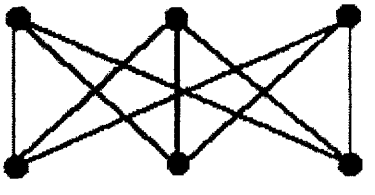
Graph theory has a long and colorful history. It started with Euler, when he studied the Königsberg problem:



Old Königsberg had seven bridges (marked *a* through *g* in the sketch). The townspeople wondered if was possible to take a walk around the town in such a way as to cross each of the seven bridges exactly once.

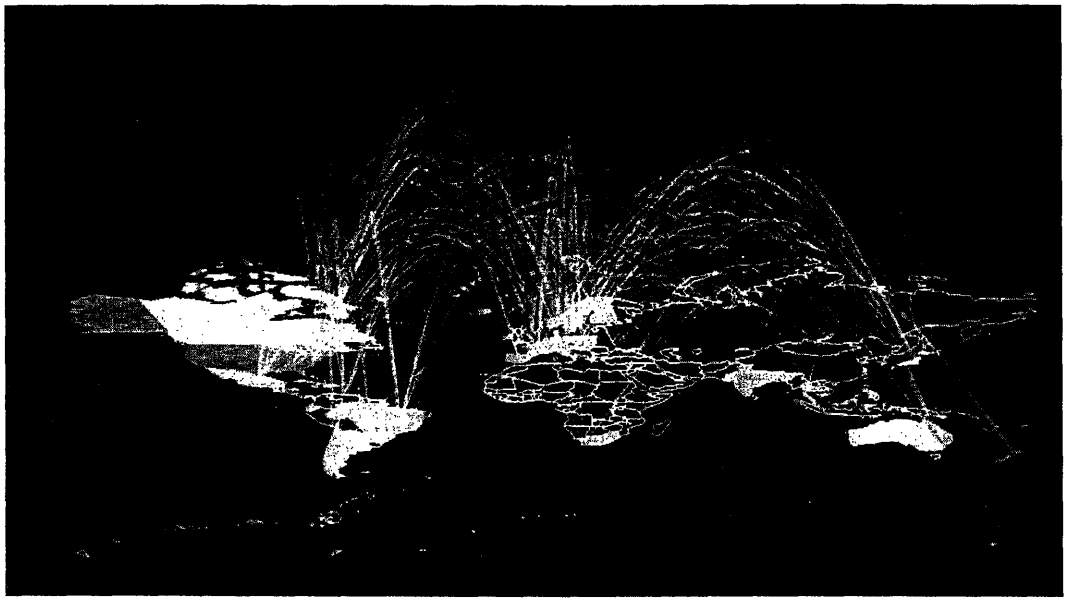
Traditional graph theory evolves around finite, regular graphs and is combinatorial in approach. Some favorite graphs include complete graphs (or cliques), bipartitie graphs, cycles, trees, grids (or lattices), and some favorite problems include Eulerian or Hamiltonian paths, chromatic numbers, and graph isomorphisms.





While traditional graph theory is well-developed and has found many diverse applications, it is clearly inadequate to deal with large, irregular, seemingly random graphs, or complex networks.

Complex, web-like structures describe a wide variety of systems spanning the spectrum from biology to internet to socialogy. For example, in a metabolic network the nodes are metabolites and the (directed) edges are chemical reactions; in a gene-regulatory network the nodes are genes, while the edges can carry both direction and weight, corresponding to activation/suppression and the strength thereof; in a protein-protein interaction network, a pair of proteins are connected if there is a physical association between them. The Internet is a complex network of routers and computers linked by various physical or wireless connections; the World Wide Web consists of individual webpages with hyperlinks both coming in and going out. In a social network, the nodes are individual persons and the edges represent various social relationships, along which ideas (or diseases!) spread and propagate.



How could one describe such complex systems? How does their network topology look like? Are there any organizing principles underlying such complex networks? How could

they have evolved? How robust are they? Can one predict their behaviors under perturbation or through evolution?

In the next a few sections we briefly overview the classical approach as well as some recent developments. We establish notations along the way, in particular introduce several key topological measures of complex networks.

3.1.2 Random Networks and the ER Model

The Hungarian mathematicians Erdös and Rényi first studied random graphs in the late 1950s, using probabilistic methods to derive large-scale, statistical properties of random graphs. In the ER model, one starts with N nodes and connects each pair of nodes with probability p, generating a random network.

For any node i, the probability that it is of degree k (i.e. connected to k other nodes in the network) follows the Binomial distribution, corresponding to k successes out of $N-1$ trials with success probability p:

$$P_i(k) = \binom{N-1}{k} p^k (1-p)^{(N-1-k)}$$

It has been show that the degree distribution of the network itself, that is, the number of nodes with a certain degree k, follows the Poisson distribution, with

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \text{ where } \lambda = P_i(k) = \binom{N-1}{k} p^k (1-p)^{(N-1-k)}.$$

Intuitively N is large and the graph is sparse, since the number of actual edges would be much smaller than $C(N,2)$ all possible edges. Hence p would be small. Furthermore,

$P_i(k)$ are close to being independent random variables. Thus to good approximation the Binomial distribution can be replaced by a Poisson.

Connectedness and Diameter

One of the greatest discoveries of ER is that many topological properties of such random graphs appear quite suddenly, at a threshold value $p(N)$, e.g. the emergence of a giant cluster. For a totally connected graph (or the single largest connected component of a disconnected graph), its diameter, or the characteristic path length, is the average distance between all pairs of nodes. One feature of random networks is that they have short typical path length. Empirically it has been observed that real networks tend to have short typical path length, close to that of comparable random networks, a phenomenon known as "six degree of separation".

Clustering Coefficient

Unlike random networks, real networks also exhibit a large degree of clustering. For example, in a social network, two acquaintences of the same person are more likely to know each other than just any two random persons. Similarly, in a protein-protein interaction network, the interaction partners of a protein are also likely to interact among themselves, since they are all involved in the same complex or process. To be precise, we define the clustering coefficient around each node as follows. Suppose a node has k neighbors and there are m edges among the neighbors. Then we define

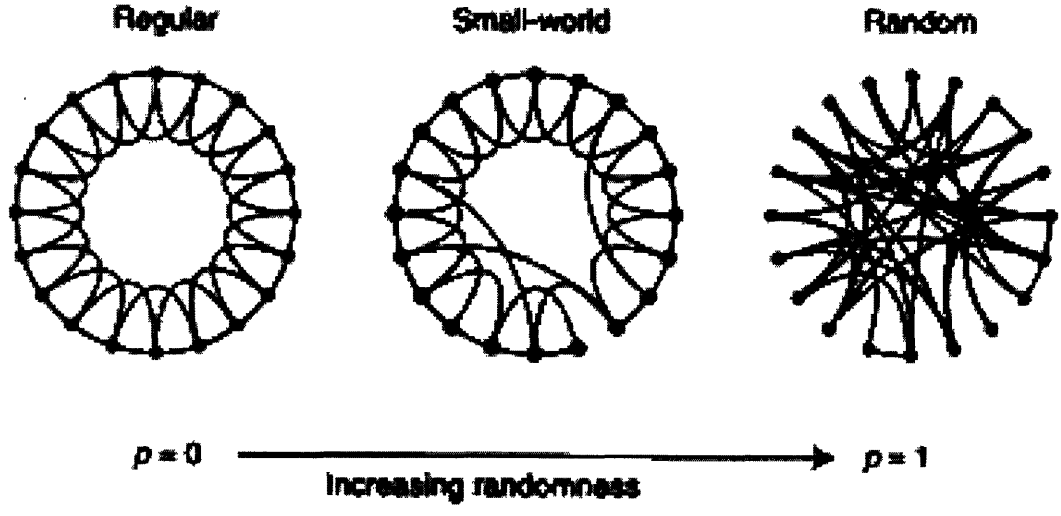
$C = \frac{m}{\binom{k}{2}}$, that is, the fraction of all possible edges that are actually present.

3.1.3 Small-world Networks and the WS Model

Up to late 1990s, the only available network models were based on either regular graphs or the ER model. Unfortunately, as data on real networks accumulate and computing

power multiplies, it has been shown that neither regular nor random graphs capture the essence of most real-world, large-scale networks.

In 1998 Watts and Strogatz [1]introduced a new model, the so-called small-world networks. A small-world network is characterized by short typical path length and high local clustering. The former property is satisfied by random graphs but not regular lattices, while the latter holds for regular lattices but not random graphs. Thus small-world networks lie between the two extremes and the WS model is constructed to describe this transition from a locally ordered system to a random network.

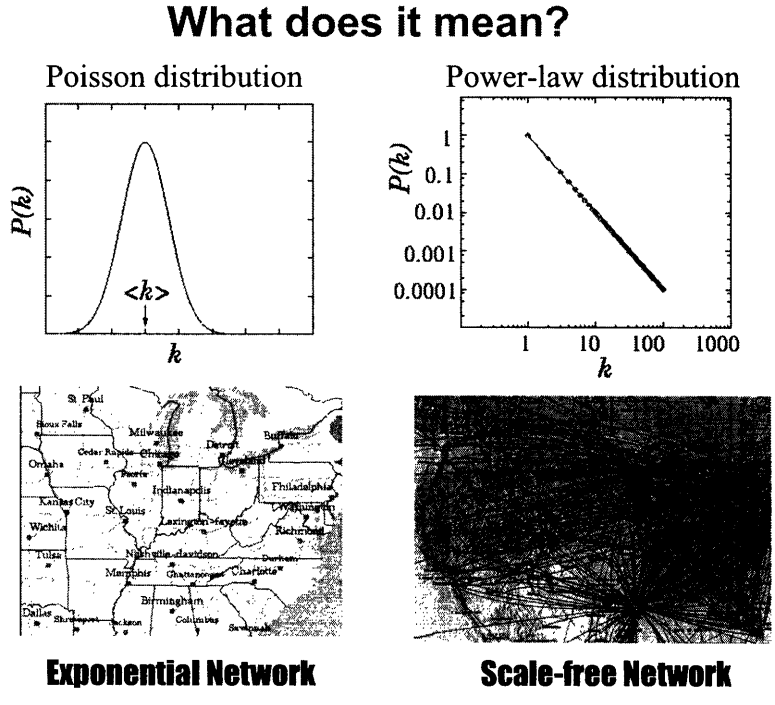


In the WS model, one starts with a regular, second-order (each node is connected to its nearest and next-nearest neighbors) ring lattice with N nodes. Then each edge is rewired with probability p, under the constraint that no two nodes can have more than one edge and no node can have an edge with itself. Thus at $p = 0$ the system is a highly clustered lattice, and the typical path length L grows linearly with the number of nodes N, while at $p = 1$ the system becomes a random graph, poorly clustered but with short typical path length $L \propto \log N$. WS has shown that in the interval $0 < p < 0.01$ the system acquires short typical path length while still highly clustered. In other words, as one introduces randomness into an orderly system, it rapidly becomes small-world. This helps to explain

why most real networks are small-world; conversely, the small-world properties have been used to validate real networks. For example, the high local clustering property has been used to assess the confidence of individual edges in a protein-protein interaction network.

3.1.4 Scale-free Networks and the BA Model

Both the ER and WS models lead to networks in which the degree distribution $P(k)$ has an exponential cutoff and is centered around $\langle k \rangle$, the average degree. However, it has been observed that many complex networks are free of scale, that is, the degree distribution decays as a power law, following $P(k) \propto k^{-\gamma}$ for some $\gamma > 0$.



One real-world example, of particular interest to scientists, is the collaboration network, where the nodes are the scientists and two nodes are connected if the two scientists have co-authored a paper together. Not surprisingly, all such networks are small-world, that is,

they exhibit short typical path length and high clustering coefficient. What is not so obvious, however, is that all such networks are also scale-free.

To understand the origin of this discrepancy between expected exponential decays and observed power-law tails, Barabási and Albert [2] have argued that two crucial aspects of real complex networks are not accounted for in the ER or WS model. First, both models have a fixed number of nodes, which are then randomly connected (ER) or reconnected (WS). In contrast, most real-world networks are continuously growing by addition of new nodes that are connected to the existing ones. For example, new webpages are being constantly created, with links to more established websites. Second, in both models the nodes are connected (or rewired) with uniform probability, while most real networks exhibit preferential attachment. For example, a new researcher is more likely to collaborate on papers with more established scientists.

Barabási and Albert have incorporated these two ingredients, missing in previous models, into a new model, which naturally leads to scale-free networks. In contrast to the ER and WS models, in which the goal is to account for network topology, the BA model focuses on network dynamics and evolution, with topology only as a byproduct.

3.1.5 Network Dynamics and Evolution

The BA model is defined in two steps:

(1) Growth: Start with a small number m_0 of nodes at time t_0. At every time step thereafter, a new node with $m \leq m_0$ edges is added to the system, that is, the new node will be connected to m existing nodes.

(2) Preferential attachment: The probability Π that a new node will be connected to existing node i depends on the degree k_i of that node, such that

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

Thus after t time steps the model leads to a random network with $N = m_0 + t$ nodes, $E = mt$ edges, and total node degree $\sum_j k_j = 2E = 2mt$.

Barabási and Albert have shown, using simulations, that such a network does evolve into a scale-invariant state, with the scaling exponent $\gamma \approx 3$, independent of m, the only parameter in the model. In addition, $P(k)$ is independent of time t (or equivalently, the system size $N = m_0 + t$), which indicates that despite its continuous growth, the system organizes itself into a scale-free stationary state. This independence of time or system size fits well with the fact that the power-law distribution holds for real complex systems of drastically different sizes and at different stages of development.

Furthermore, Barabási and Albert were able to derive $P(k)$ analytically, using a mean field approach to calculate the time evolution of the degree of a given node. Consider node i with degree $k_i(t)$. Since at each time step m edges are added, each connecting to node i with probability $\Pi(k_i)$, the rate of change of k_i is just

$$\frac{\partial k_i}{\partial t} = m\Pi(k_i) = m\frac{k_i}{\sum_j k_j} = m\frac{k_i}{2mt} = \frac{k_i}{2t}$$

The solution of this first-order differential equation, with the initial condition that node i was added to the system at time t_i with degree $k_i(t_i) = m$, is

$$k_i(t) = m\sqrt{\frac{t}{t_i}}$$

Thus all nodes acquire more edges over time, while the older nodes (those with smaller t_i) increase their degrees faster.

To calculate $P(k)$, we have

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{\partial}{\partial k} P\left(m\sqrt{\frac{t}{t_i}} < k \right) = \frac{\partial}{\partial k} P\left(t_i > \frac{m^2 t}{k^2} \right) = 1 - \frac{\partial}{\partial k} P\left(t_i \le \frac{m^2 t}{k^2} \right), \quad (1)$$

Since we picked node i at random out of $N = m_0 + t$ nodes, $P(t_i)$ follows the Uniform distribution with "height" $\frac{1}{m_0 + t}$. Thus $P\left(t_i \le \frac{m^2 t}{k^2} \right) = \frac{m^2 t}{k^2 (m_0 + t)}$, (2)

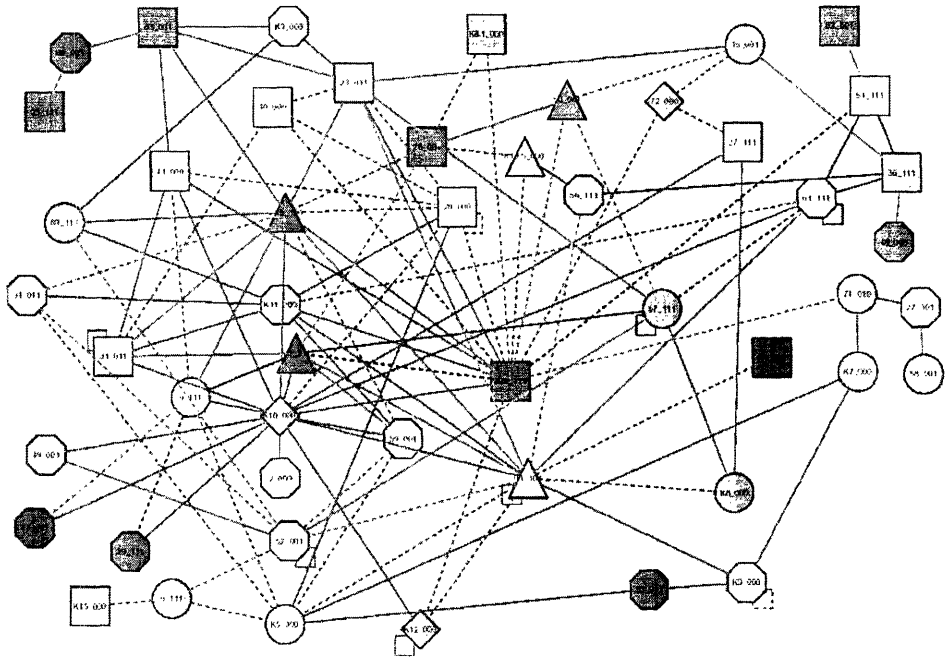
Combining (1) and (2), we obtain

$$P(k) = \frac{2m^2 t}{m_0 + t} k^{-3}$$

Thus the BA model naturally leads to scale-free networks with power coefficient $\gamma = 3$. Furthermore, they have shown that both growth and preferential attachment are essential to the evolution of scale-free networks.

After BA, various new models, incorporating different "real-world" features, have been proposed. For example, the preferential attachment function was allowed to be non-linear, internal edges were allowed to be inserted or deleted within the existing system, nodes were allowed to duplicate themselves or "retire" or have different levels of "fitness". It is worth noticing, however, that all current dynamic network evolution models lead to $\gamma \ge 1$.

3.2 The KSHV protein-protein interaction network

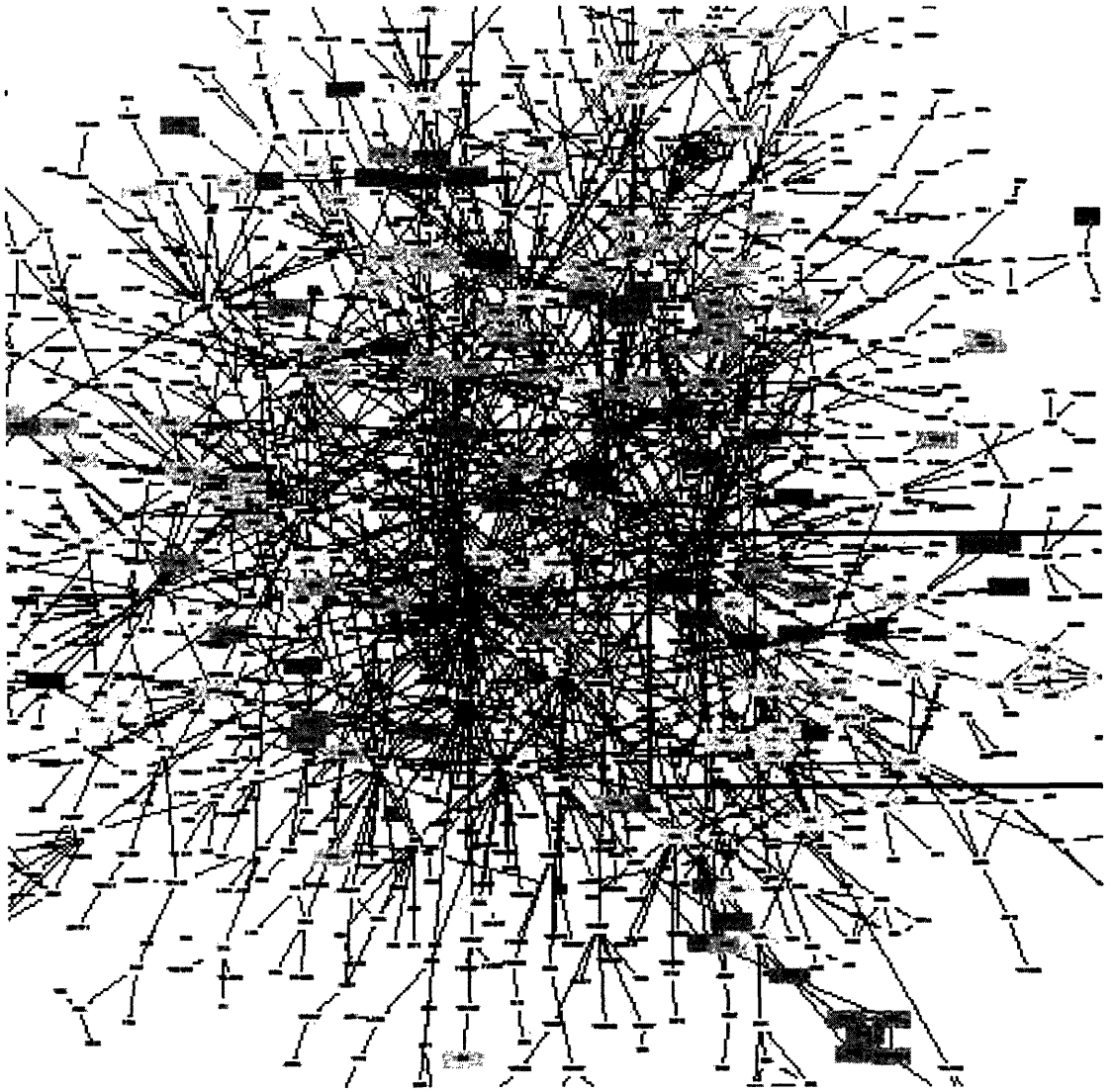


Thus far we have mostly treated the 123 interactions (115 when dimers are removed) among 50 KSHV proteins as a binary dataset. However, there are many advantages in viewing them as a network. For example, unexpected links between different complexes or cellular processes might emerge, the confidence of individual interactions could be assessed, and functional assignment could be made for proteins of unknown function. Furthermore, large-scale, system-level properties are only available through this network point of view.

We connected the binary KSHV interactions into a single network, which constitutes the first meaningful viral system. Each node represents an interacting KSHV protein; its phylogenetic class is given after the ORF name; node color represents its functional class, while node shape corresponds to its expression class. The solid edges represent

interactions confirmed by CoIP, while those dashed edges correspond to interactions positive only in the Y2H screens.

It has been shown that most real-world complex networks are both small-world and scale-free. In particular, all known biological networks are scale-free, with power coefficient $\gamma > 2$. Moreover, all current dynamic network evolution models predict $\gamma \geq 1$. Since the KSHV network is the first viral system available, we wondered if its network properties are similar to those of cellular systems and hence confirm the universality of those properties in all kingdoms of life, or are distinct. At first sight, due to the abundance of hubs and the interactions among them, we thought the KSHV network is not scale-free. Now we investigate the major network properties of KSHV and compare it to a high-confidence yeast protein-protein network, a prototypical cellular system.

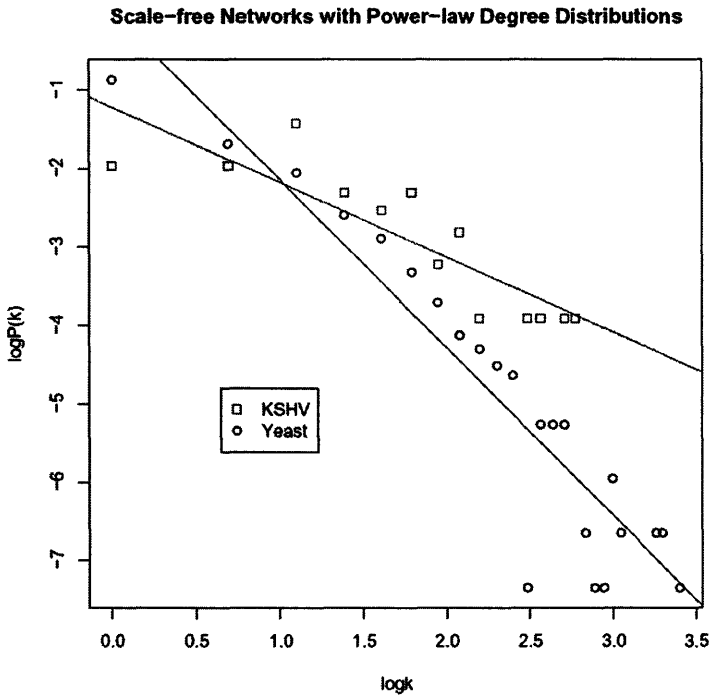


Schwikowski et al has combined Y2H and biochemical protein-protein interaction data in public databases into a single, high-confidence yeast network. Shown is the single largest connected component with 2358 edges among 1548 nodes.

3.3 Degree Distribution and Attack Tolerance

While the yeast network is typically scale-free with a power coefficient $\gamma = 2.14 \left(p = 3.64 \times 10^{-11}\right)$, the viral system has a surprisingly small scaling exponent $\gamma = 0.95 \left(p = 1.24 \times 10^{-4}\right)$. Thus albeit scale-free, the KSHV network is distinct from all

previously known biological networks, which all have $\gamma > 2$, and it cannot be explained by all current dynamic network evolution models, which all predict $\gamma \geq 1$.

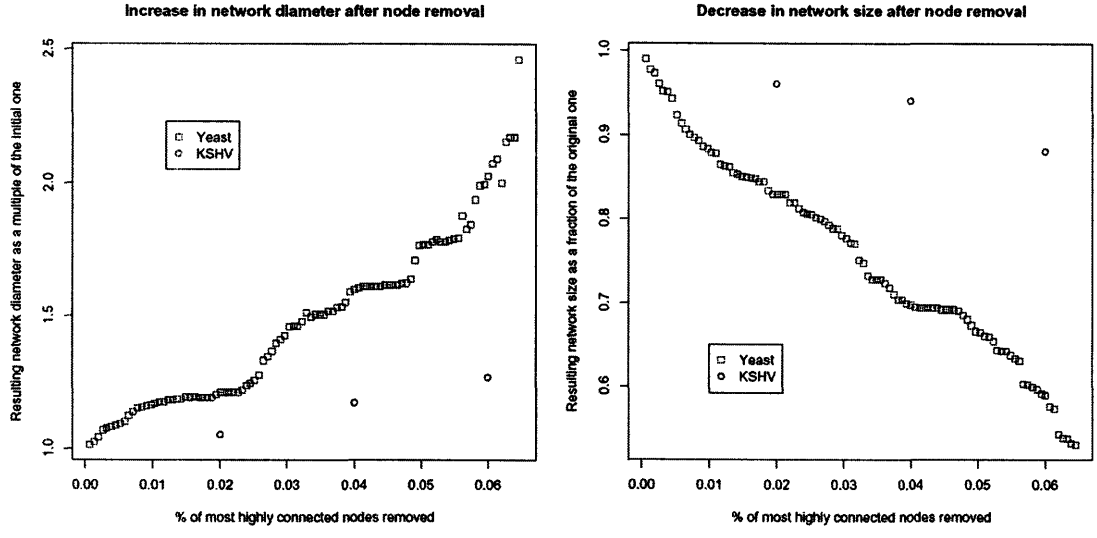


The degree distributions of the KSHV and yeast networks are plotted on a log-log scale. For each network, the probability $P(k)$ of a given node degree k is defined as the relative frequency of such nodes, that is, the number of nodes with degree k divided by the total number of nodes in the network. Then the power coefficient γ is estimated through regression analysis. For each network, both the original data points and the fitted regression line are shown.

As we now demonstrate, this unusual topological feature of KSHV has important consequences, in particular it leads to increased attack tolerance [16].

Scale-free networks are highly resistent towards random failure, but highly vulnerable under deliberate attack. Compared to yeast network,

KSHV Network Has Much Higher Attack Tolerance



To consider the robustness of a network, we consider how the network topology changes under node failures. When a node fails, we take out that node and all edges associated with it. To evaluate the remaining network, we look at the size (number of nodes) and the characteristic path length of its single largest connected component.

It has been shown that scale-free networks are highly robust against random node failures, while highly vulnerable under deliberate attack, where nodes with the highest degrees are in turn removed. This corresponds well to real-world networks, where components fail at random all the time without bringing down the whole network, but bringing down central hubs would bring down the network as a whole.

Now we demonstrate that the KSHV network has much higher attack tolerance than the yeast network. In KSHV, the top three most highly connected nodes were removed, which corresponds to 6% node removal. Similarly, we remove the top 6% most highly

connected nodes in the yeast network. As shown in the panel of figures, the KSHV network is much more stable under attack, as measured by either network size or characteristic path length.

3.4 Degree Correlation and Modularity

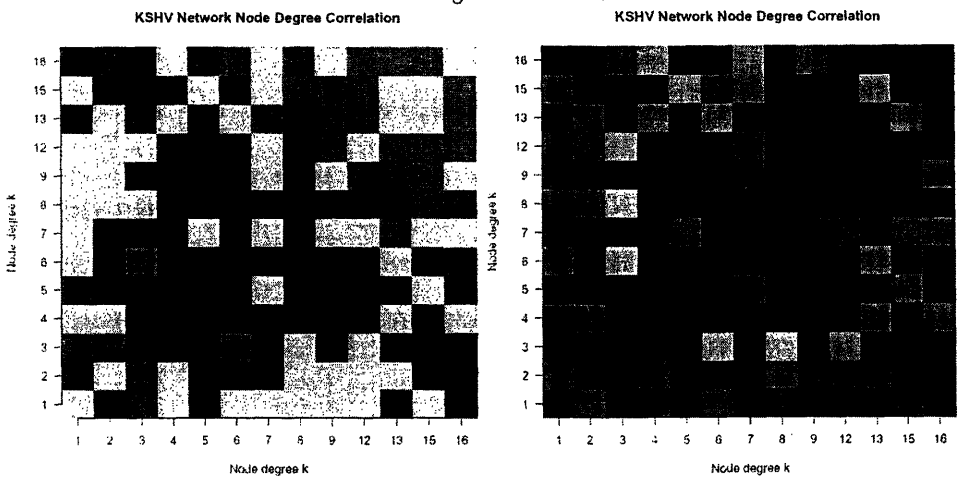
In this section we introduce a new topological feature of networks, namely the correlation of node degrees for all the interacting nodes in the network. Along the way, we also introduce a new network randomization technique, which will be used in subsequent sections.

Maslov and Sneppen [17] have shown that, for the yeast network, hubs tend to avoid each other and connect to those low-degree nodes. Furthermore, Berg and Lässig [18]have shown, using statistical mechanics, that this degree correlation can be accomplished with a properly chosen partition function.

To estimate the statistical significance of observed degree correlation in a real network, one must generate an ensemble of "comparable" random networks. A simple choice would be networks generated by the ER model, with the same number of nodes and edges (that is, connection probability p) as the real network. However, such networks lack a key property of the real network: while the real network is scale-free with a power-law degree distribution, the simulated networks have a Poisson degree distribution. To ensure the observed degree correlation is not an artifact of network topology, Maslov and Sneppen constructed random networks with the same degree distribution, using an edge-swapping strategy. In such a randomzied network, each node has the same number of edges as before; a pair of edges are picked at random and swapped, provide the swapping does not create redundant edges or self loops; after a number of such swappings one obtains a random network with exactly the same degree distribution – each node has the same number of interaction partners as before and only the identity of them are different.

It has been shown for yeast network (Maslov and Sneppen, Science 2002) that hubs tend to avoid each other and connect to low-degree nodes. We do not observe such a clear pattern for

KSHV Degree Correlation



Random Model (logodds) Maslov-Sneppen Model (z-scores)

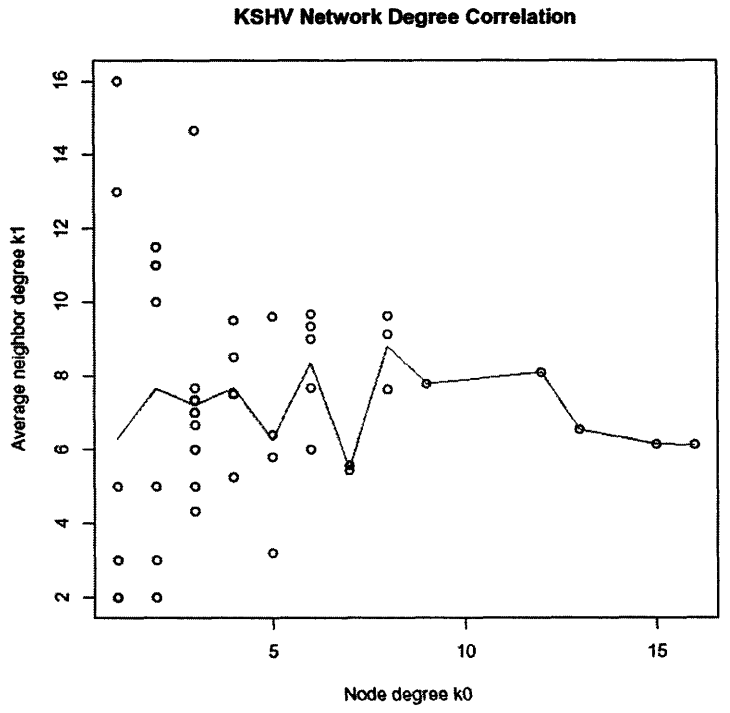
For each pair of node degrees (k_i, k_j), we find its observed frequency in the real network, that is, we increase the counter by 1 if there is an edge connecting a pair of nodes with degree k_i and k_j. To estimate the statistical significance of the observed degree correlations, we compute two separate measures, logodds and (empirical) z-scores.

For each pair (k_i, k_j), its observed frequency is $P_{ob}(k_i, k_j) = \#(k_i, k_j)/E$, where $\#(k_i, k_j)$ is the number of edges connecting node degrees k_i and k_j, and E is the number of edges in the network.

To compute empirical z-scores, we follow the MS procedure to generate 1000 random networks with the same degree distribution as the real network. Then the count for each degree pair (k_i, k_j) can be compared to those in random networks and z-scores derived.

In both heat maps, the square at position (i, j) corresponds to the degree correlation between node degree k_i and k_j, while the color intensity corresponds to the amplitude of

the statistics, with green denoting suppression and red denoting enrichment. For example, bright red squares are those with big positive logodds or z-scores, bright green squares are those with big negative logodds or z-scores, while those dark squares correspond to those degree correlations close to random, with logodds or z-scores close to 0.



For a different view, we plotted the average degree of its neighbors as a function of the node degree itself. Again, we do not observe a declining degree correlation for KSHV.

3.5 Low Clustering and Dynamic Mode of Action

3.5.1 Characteristic Path Length

Both the KSHV and yeast networks exhibit short typical path length, comparable to random networks of the same size. For comparison, we have generated random networks under both the ER and the MS model and computed their characteristic path length.

	L	L_{ER}	L_{MS}
KSHV	2.84	2.68	2.60
Yeast	7.28	6.43	6.35

3.5.2 Clustering Coefficient

For each of the KSHV and yeast networks, we compute its clustering coefficient C and compare it to that of random networks under either ER or MS model. Under the ER model, for a network with E edges among N nodes, the clustering coefficient is a constant, namely $C_{ER} = E \Big/ \binom{N}{2}$, since the edges are placed uniformly at random. For the MS model, we generate 1000 random networks using the edge-swapping strategy, and define C_{MS} to be the average clustering coefficient of those random networks.

We have the following results:

	C	C_{ER}	C_{MS}
KSHV	0.146	0.094	0.193
Yeast	0.213	0.002	0.008

Thus under both models, the clustering coefficient of KSHV is comparable to those of random networks, hence the KSHV system is not small-world! In contrast, the clustering coefficient of the yeast network corresponds to about 25-100 fold enrichment over comparable random networks.

One major use of protein-protein interaction networks is to discover functional modules by locating locally dense neighborhoods, in particular cliques. Since the KSHV network does not exhibit the high local clustering property of small-world networks, we would like to explore its implications in terms of finding cliques. Since the KSHV network is relative small, we can enumerate cliques of all orders recursively, and compute their

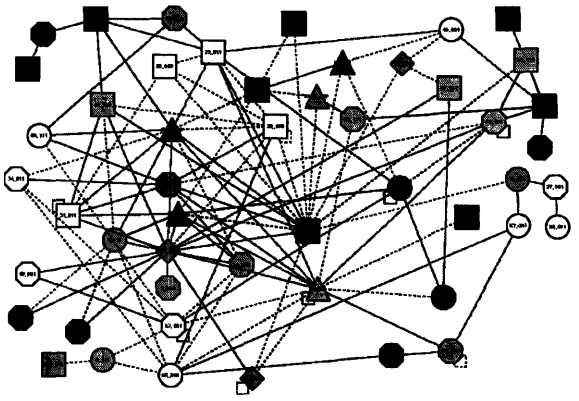
statistical significance following standard procedure. The following table summarizes the results:

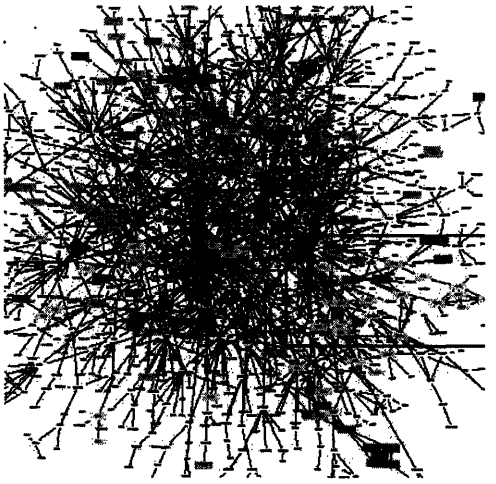
Cliques	Mean	SD	z-score	p-value	ee-value
['ORF 28', 'ORF 29b', 'ORF 41', 'ORF K10']	0.502	0.181	2.740	0.006	0.138
['ORF 28', 'ORF 29b', 'ORF 67.5', 'ORF K10']	0.603	0.181	2.179	0.034	0.137
['ORF 23', 'ORF 28', 'ORF 29b', 'ORF 67.5']	0.510	0.187	2.611	0.009	0.039
['ORF 23', 'ORF 28', 'ORF 29b', 'ORF 30']	0.369	0.173	3.633	0.000	0.014
['ORF 28', 'ORF 41', 'ORF K10']	0.428	0.269	2.122	0.061	1.224
['ORF 29b', 'ORF 41', 'ORF K10']	0.557	0.264	1.672	0.141	0.706
['ORF 31', 'ORF 41', 'ORF K10']	0.381	0.262	2.351	0.038	0.386
['ORF 41', 'ORF 9', 'ORF K10']	0.405	0.264	2.250	0.045	0.452
['ORF 31', 'ORF 67.5', 'ORF K10']	0.512	0.268	1.816	0.106	0.532
['ORF 31', 'ORF 68', 'ORF K10']	0.426	0.269	2.132	0.061	1.226
['ORF 28', 'ORF 67.5', 'ORF K10']	0.564	0.273	1.594	0.160	0.643
['ORF 29b', 'ORF 67.5', 'ORF K10']	0.697	0.256	1.182	0.324	0.324
['ORF 59', 'ORF 67.5', 'ORF K10']	0.509	0.268	1.825	0.105	0.528
['ORF 60', 'ORF 67.5', 'ORF K10']	0.667	0.259	1.281	0.277	0.277
['ORF 67.5', 'ORF 9', 'ORF K10']	0.541	0.269	1.700	0.132	0.265
['ORF 29b', 'ORF 68', 'ORF K10']	0.607	0.268	1.460	0.203	0.815
['ORF 59', 'ORF 68', 'ORF K10']	0.423	0.268	2.148	0.056	1.128
['ORF 60', 'ORF 68', 'ORF K10']	0.573	0.272	1.567	0.169	0.677
['ORF 68', 'ORF 9', 'ORF K10']	0.446	0.266	2.073	0.065	0.525
['ORF 39', 'ORF 9', 'ORF K10']	0.289	0.233	3.046	0.007	0.103
['ORF 47', 'ORF 9', 'ORF K10']	0.290	0.234	3.020	0.007	0.107
['ORF 29b', 'ORF K10', 'ORF K12']	0.452	0.250	2.189	0.057	0.630
['ORF 60', 'ORF K10', 'ORF K12']	0.425	0.254	2.263	0.047	0.518
['ORF 23', 'ORF 63', 'ORF K9']	0.167	0.205	4.052	0.002	0.069
['ORF 23', 'ORF 28', 'ORF 67.5']	0.435	0.275	2.045	0.071	0.284
['ORF 23', 'ORF 29b', 'ORF 67.5']	0.571	0.274	1.561	0.169	0.169
['ORF 23', 'ORF 60', 'ORF 67.5']	0.540	0.274	1.675	0.137	0.137
['ORF 23', 'ORF 63', 'ORF 67.5']	0.352	0.259	2.496	0.030	0.091
['ORF 28', 'ORF 29b', 'ORF K11']	0.459	0.274	1.970	0.082	0.492
['ORF 28', 'ORF 29b', 'ORF K10']	0.608	0.267	1.464	0.205	0.822
['ORF 28', 'ORF 29b', 'ORF 41']	0.415	0.267	2.183	0.051	1.038
['ORF 28', 'ORF 29b', 'ORF 67.5']	0.545	0.274	1.655	0.143	0.572
['ORF 23', 'ORF 28', 'ORF 29b']	0.490	0.274	1.855	0.098	0.394
['ORF 23', 'ORF 29b', 'ORF 30']	0.376	0.253	2.458	0.032	0.196
['ORF 36', 'ORF 54', 'ORF 61']	0.141	0.195	4.382	0.001	0.117
['ORF 60', 'ORF 61', 'ORF K10']	0.523	0.268	1.776	0.117	0.589
['ORF 60', 'ORF 61', 'ORF K11']	0.377	0.265	2.346	0.038	0.774
['ORF 23', 'ORF 28', 'ORF 30']	0.255	0.238	3.118	0.009	0.218
['ORF 28', 'ORF 29b', 'ORF 30']	0.355	0.254	2.529	0.024	0.597

Thus the low clustering coefficient translates into the lack of distinct complexes and functional modules in the KSHV network – not only are higher-order cliques absent, but the number of cliques at each level is not enriched.

3.6 Summary of Results

Comparision of Network Properties of KSHV and Yeast





	N	E	<k>	r	L/Lran/Lms	C/Cran/Cms
KSHV	50	115	4.60	0.95	2.84/2.68/2.60	0.15/0.09/0.19
Yeast	1548	2358	3.05	2.14	7.28/6.43/6.35	0.21/0.00/0.01

Albeit scale-free, KSHV network differs in many aspects from all known networks.

For comparison, we put the KSHV and yeast networks side by side, together with important topological quantities. The unusual features of KSHV are highlighted in red.

3.7 Discovering Functional Associations through Interaction Patterns

Due to the huge amount of data accumulated by various genome sequencing and genome-scale experiments, it is increasingly important to transfer our existing knowledge to new data and new systems. In the past, functional associations, in particular protein-protein interactions, have been predicted on the basis of similar functional classes, phylogenetic classes, correlated expression profiles, shared cellular compartments, local clustering in networks, or a combination of those approaches. In this section we explore new ways to discover functional associations based on topological properties, and propose some predictions for experimental verification.

3.7.1 Neighbors in Common

While the clustering coefficient around each node is a local measure, namely it represents the likelihood of interaction among the node's neighbors, interaction is not strictly a local event. A physical interaction involve two proteins, and the effect extends to at least the two nodes's other interaction partners. Furthermore, there are many cases of biological significance where two proteins do not directly interact, but are bridged by a third protein. Clearly the local measure of clustering coefficient does not adequately capture this second-order effect. We now introduce a new measure, which addresses both of the concerns at once.

Given any two nodes in the network at either distance 1 (i.e. directly interacting) or distance 2 (i.e. bridged by one other node), we look at the number of neighbors they have in common. The idea behind this is as follows – if two nodes directly interact and they have common neighbors, then the confidence of this interaction is enhenced; if two nodes do not directly interact but are bridged by many common neighbors, then they are likely to be functionally related.

[orf1 orf2]	deg1-1	deg2-1	#(common neighbors)
['ORF 2' 'ORF K10']	0	15	0
['ORF 23' 'ORF 45']	8	3	0
['ORF 25' 'ORF 65']	0	1	0
['ORF 27' 'ORF 74']	1	2	0
['ORF 28' 'ORF K5']	7	6	0
['ORF 29b' 'ORF 50']	14	3	0
['ORF 29b' 'ORF 54']	14	2	0
['ORF 29b' 'ORF 72']	14	2	0
['ORF 29b' 'ORF 74']	14	2	0
['ORF 29b' 'ORF K10.5']	14	2	0
['ORF 29b' 'ORF K8.1']	14	1	0
['ORF 31' 'ORF 30']	5	3	0
['ORF 31' 'ORF K11']	5	7	0
['ORF 34' 'ORF 52']	3	6	0
['ORF 34' 'ORF 67.5']	3	11	0
['ORF 34' 'ORF K11']	3	7	0
['ORF 34' 'ORF K5']	3	6	0
['ORF 36' 'ORF 45']	4	3	0
['ORF 36' 'ORF 48']	4	0	0
['ORF 36' 'ORF 54']	4	2	0
['ORF 36' 'ORF 61']	4	4	0
['ORF 37' 'ORF 72']	2	2	0
['ORF 37' 'ORF K10']	2	15	0
['ORF 37' 'ORF K8']	2	2	0
['ORF 45' 'ORF 50']	3	3	0
['ORF 45' 'ORF 72']	3	2	0
['ORF 49' 'ORF 52']	1	6	0
['ORF 49' 'ORF K10']	1	15	0
['ORF 53' 'ORF K3']	1	2	0
['ORF 53' 'ORF K5']	1	6	0
['ORF 54' 'ORF 62']	2	0	0
['ORF 56' 'ORF 36']	2	4	0
['ORF 56' 'ORF K10.5']	2	2	0
['ORF 57' 'ORF 23']	5	8	0
['ORF 57' 'ORF 50']	5	3	0
['ORF 57' 'ORF 52']	5	6	0

['ORF 57' 'ORF 61']	5	4	0
['ORF 57' 'ORF 68']	5	7	0
['ORF 57' 'ORF K8']	5	2	0
['ORF 58' 'ORF 27']	0	1	0
['ORF 59' 'ORF 52']	5	6	0
['ORF 59' 'ORF K11']	5	7	0
['ORF 59' 'ORF K5']	5	6	0
['ORF 6' 'ORF 52']	2	6	0
['ORF 6' 'ORF K15']	2	0	0
['ORF 6' 'ORF K5']	2	6	0
['ORF 60' 'ORF 52']	12	6	0
['ORF 60' 'ORF 56']	12	2	0
['ORF 60' 'ORF K1']	12	0	0
['ORF 60' 'ORF K3']	12	2	0
['ORF 60' 'ORF K5']	12	6	0
['ORF 60' 'ORF K8']	12	2	0
['ORF 63' 'ORF 41']	4	5	0
['ORF 63' 'ORF 65']	4	1	0
['ORF 69' 'ORF 52']	3	6	0
['ORF 69' 'ORF 67.5']	3	11	0
['ORF 69' 'ORF K11']	3	7	0
['ORF 69' 'ORF K9']	3	2	0
['ORF 75' 'ORF 50']	4	3	0
['ORF 75' 'ORF 67.5']	4	11	0
['ORF 75' 'ORF 68']	4	7	0
['ORF 75' 'ORF K10.5']	4	2	0
['ORF 75' 'ORF K8.1']	4	1	0
['ORF K7' 'ORF 74']	2	2	0
['ORF K7' 'ORF K3']	2	2	0
['ORF K7' 'ORF K5']	2	6	0
['ORF 23' 'ORF K9']	8	2	1
['ORF 28' 'ORF K11']	7	7	1
['ORF 29b' 'ORF 68']	14	7	1
['ORF 29b' 'ORF K11']	14	7	1
['ORF 29b' 'ORF K12']	14	2	1
['ORF 31' 'ORF 41']	5	5	1
['ORF 31' 'ORF 67.5']	5	11	1

['ORF 31' 'ORF 68']	5	7	1
['ORF 59' 'ORF 67.5']	5	11	1
['ORF 59' 'ORF 68']	5	7	1
['ORF 60' 'ORF 23']	12	8	1
['ORF 60' 'ORF 68']	12	7	1
['ORF 60' 'ORF K11']	12	7	1
['ORF 60' 'ORF K12']	12	2	1
['ORF 61' 'ORF K10']	4	15	1
['ORF 61' 'ORF K11']	4	7	1
['ORF 63' 'ORF 67.5']	4	11	1
['ORF 63' 'ORF K9']	4	2	1
['ORF 9' 'ORF 39']	5	1	1
['ORF 9' 'ORF 41']	5	5	1
['ORF 9' 'ORF 47']	5	1	1
['ORF 9' 'ORF 67.5']	5	11	1
['ORF 9' 'ORF 68']	5	7	1
['ORF K10' 'ORF 39']	15	1	1
['ORF K10' 'ORF 47']	15	1	1
['ORF 23' 'ORF 30']	8	3	2
['ORF 28' 'ORF 30']	7	3	2
['ORF 28' 'ORF 41']	7	5	2
['ORF 29b' 'ORF 30']	14	3	2
['ORF 29b' 'ORF 41']	14	5	2
['ORF 59' 'ORF K10']	5	15	2
['ORF 60' 'ORF 61']	12	4	2
['ORF 60' 'ORF 67.5']	12	11	2
['ORF 63' 'ORF 23']	4	8	2
['ORF K12' 'ORF K10']	2	15	2
['ORF 23' 'ORF 28']	8	7	3
['ORF 28' 'ORF 67.5']	7	11	3
['ORF 28' 'ORF K10']	7	15	3
['ORF 29b' 'ORF 23']	14	8	3
['ORF 29b' 'ORF 67.5']	14	11	3
['ORF K10' 'ORF 31']	15	5	3
['ORF 23' 'ORF 67.5']	8	11	4
['ORF 60' 'ORF K10']	12	15	4
['ORF K10' 'ORF 41']	15	5	4

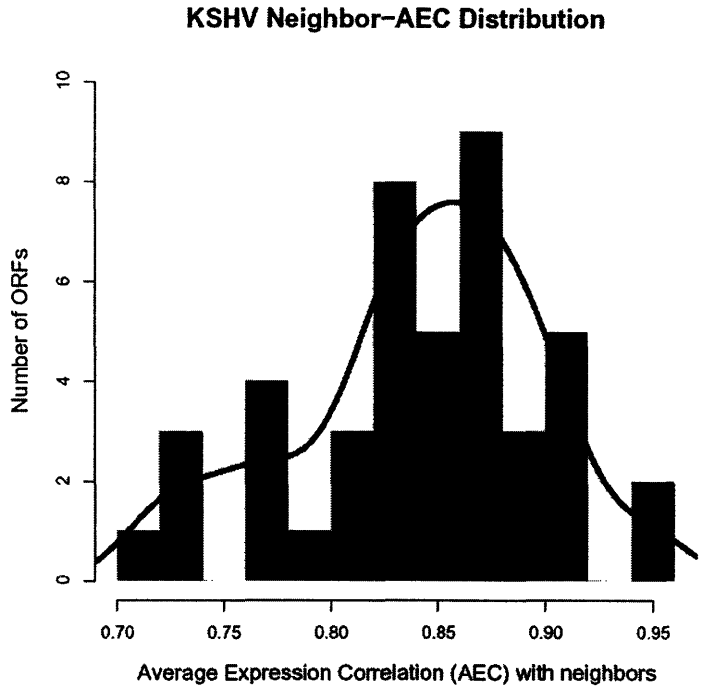
['ORF 29b'□ 'ORF K10']	14	15	5
['ORF 9'□ 'ORF K10']	5	15	5
['ORF K10'□ 'ORF 68']	15	7	5
['ORF 29b'□ 'ORF 28']	14	7	6
['ORF K10'□ 'ORF 67.5']	15	11	6

The table summarizes the result for 115 non-dimeric KSHV interactions.

Note that most interacting proteins have no or few neighbors in common, a consequence of the lack of local clustering in the KSHV network, and hence it is all the more significant for those interacting proteins sharing many common neighbors, e.g. K10-67.5, K10-68, 23-67.5, 29b-28, 9-K10, K10-41, 60-K10, 29b-K10. The most extreme example would be 9-K10, where K10 has all the interactions ORF 9 has! ORF 9 is DNA polymerase, one of the most well studied proteins, while the role of ORF K10 is still under investigation. Our result suggests that K10 is also implicated in DNA replication.

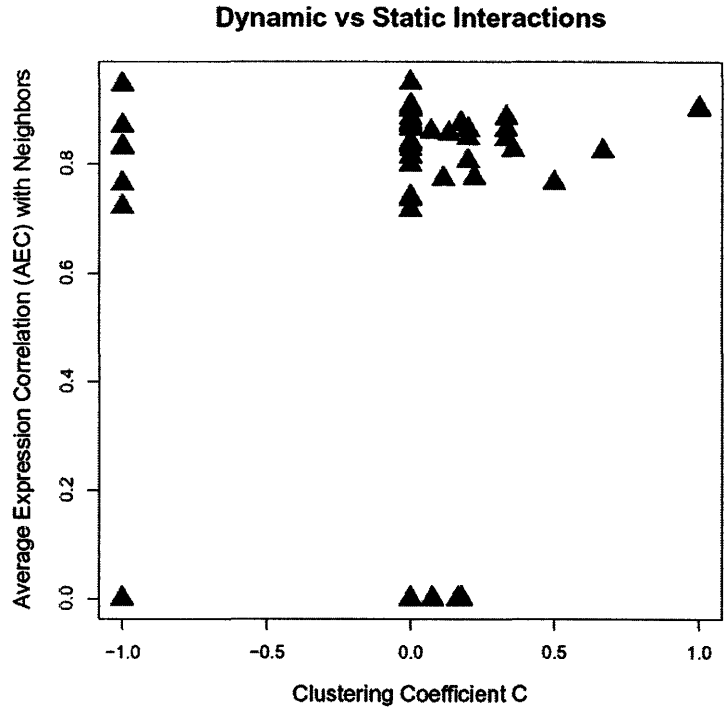
3.7.2 Clustering Coefficient with Average Expression Correlation

To account for the low local clustering observed for the KSHV network, we argue that many of the interactions are dynamic rather than static. While all protein-protein interactions are connected into a single network, one has to keep in mind that this is a superimposed view – the interactions could take place at different time or place or under different conditions. The clustering coefficient around each node addresses the space constraint. To take into account the time constraint, we introduce a new measure, the average expression correlation (AEC) around each node. For any given node with an expression profile, we look at each of its neighbors which also have an expression profile in turn and compute the correlation between the two profiles. The AEC around that node is defined as the average of those correlations. Thus the AEC around a node measures how similarly that node and its neighbors are expressed.



Even with limited data, AEC does distribute nicely.

Now we consider the clustering coefficient (C) and the average expression correlation (AEC) with neighbors around each node. If a node's interactions are largely static, as in a complex, then 1) its neighbors are more likely to interact with each other (hence high C), and 2) it is more likely to be similarly expressed as its neighbors (hence high AEC). On the other hand, if the interactions take place at different time/place, then AEC/C would be lower. So, we combine the two measures and use them to classify KSHV nodes in the C-AEC space – different regions on the plane would then correspond to different modes of action and nodes clustered together would then have putative functional associations.



The distribution of 50 KSHV nodes on the C-AEC plane. C is set to -1 if a node has only one interaction partner (and hence C is undefined), AEC is set to 0 if no pairwise expression profiles are available between a node and any of its neighbors.

1) All hubs are not the same. For example, K10 and 29b are the top two hubs, but K10 has much higher C and AEC. (AEC for 29b is rather low -- one would hypothesize that it interacts with its partners at different time.)

2) In the manuscript you hypothesized about the four IE proteins 50, 57, K8, and 45. Based on C (all four =0.0) and AEC, it would seem unlikely that they form a complex.

3) ORFs 36 and 54 have very low AEC (ORF 36 has the lowest AEC among all KSHV nodes) and C (both have C=0) – this is not too surprising (actually quite reassuring), since both are enzymes.

4) ORFs 9 and 41 have high C and AEC – this confirms their roles in the DNA replication machinery. But they have only one neighbor in common (data from analysis in the previous section, not shown), which suggests they are not in the same complex.

5) K3, K5, and K7 all have very low C but very high AEC – do they have similar biological roles, since they share similar action patterns? Their putative functional association discovered by our analysis already has biological support – the "K" in their ORF names stands for KSHV-specific, so they indeed belong to the same group of genes.

6) K5 and 52 share the following feature – each has 7 neighbors and there is no interaction among any of the neighbors – are they involved in diverse roles? Moreover, K5 and 52 have 4 neighbors in common – are they functionally or structurally similar?

3.8 *Joint Analysis using C and AEC for Yeast*

To both validate our methodology and obtain new results, we now apply the joint C_AEC analysis to a prototypical cellular system, the single largest connected component (SLCC) of the yeast PPI network from Benno et al, which has 2358 edges among 1548 nodes.

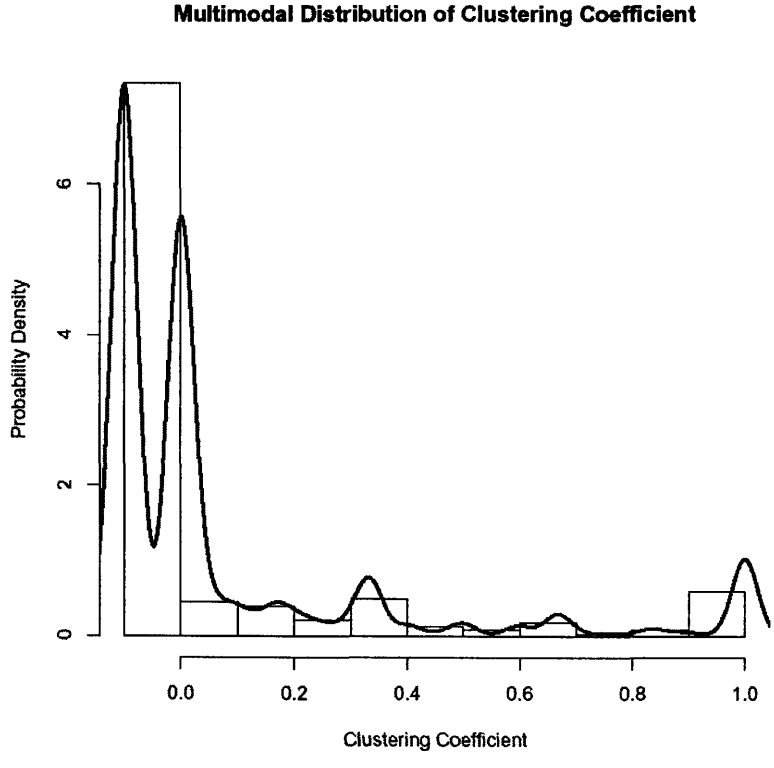
Unlike KSHV, for which there has been only one genome-wide expression profiling analysis to date [19], there have been several such studies on yeast, under diverse conditions. Here we use the cell cycle dataset from Spellman et al [20] as a representative, in which 6178 genes are profiled under 77 distinct conditions or time points.

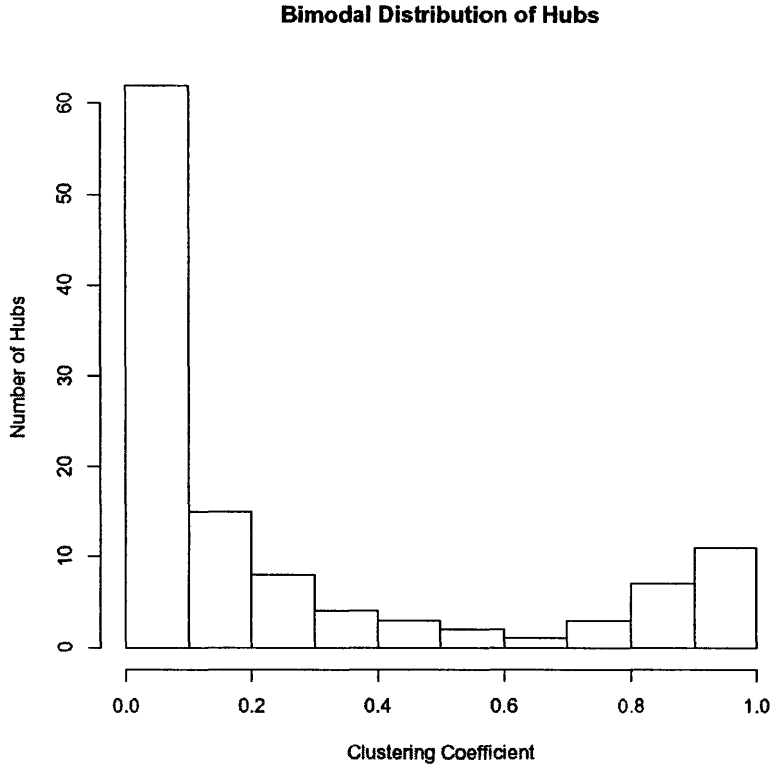
After removing 12 nodes in the SLCC that do not have expressions, we obtain a further SLCC with 2333 edges among 1531 nodes. All further analysis is done on this final SLCC. Furthermore, to strengthen data, we also consider hubs separately, since C and AEC around low-connectivity nodes tend to fluctuate a lot due to the small number of observations. Among the 1531 nodes, we have 116 hubs, defined as those nodes with 8 or more neighbors. Now we look at C, AEC, and their combination in turn with respect to this SLCC and its core set of hubs.

For an independent biological validation of our analysis, we use the Gene Ontology (GO) annotations available from SGD. For each ORF, we extract its biological process, biochemical function, and cellular component information.

3.8.1 Clustering Coefficient

We first compute C around each node for the entire SLCC and plot this background distribution. For those nodes with only one neighbor and hence C is undefined, we set C equal to -0.1. While the peaks at -0.1 and 0 are expected, there are also distinct peaks at over 0.3 and close to 1, suggesting typical interaction patterns within biological modules.





Strikingly, C for hubs exhibits a bimodal distribution at both ends, that is, hubs tend to have either very low C or very high C, suggesting drastically different modes of action. For example, presumably those hubs on the left end are enzymes while those on the right lie in permanent complexes. To test this hypothesis, we look at the GO annotations for the 18 hubs with $C > 0.8$ and 62 hubs with $C < 0.1$.

Hubs with High C

YBL084C	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YBR081C	protein complex assembly*	structural molecule activity	mitochondrion*
YBR198C	transcription initiation from Pol II promoter*	general RNA polymerase II transcription factor activity	SAGA complex*
YDL008W	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YDR118W	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YDR145W	transcription initiation from Pol II promoter*	general RNA polymerase II transcription factor activity	SAGA complex*

YFR036W	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YGL112C	transcription initiation from Pol II promoter*	general RNA polymerase II transcription factor activity	SAGA complex*
YGL240W	ubiquitin-dependent protein catabolism*	enzyme regulator activity	mitochondrion*
YHR099W	regulation of transcription from Pol II promoter*	histone acetyltransferase activity	histone acetyltransferase complex*
YHR166C	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YKL022C	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YLR055C	histone acetylation*	transcription cofactor activity	nucleus*
YLR102C	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YNL172W	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YOL148C	histone acetylation*	transcription cofactor activity	SAGA complex*
YOR249C	ubiquitin-dependent protein catabolism*	protein binding*	anaphase-promoting complex
YPL254W	transcription from Pol II promoter*	transcription cofactor activity	SAGA complex*

Indeed those hubs with high C are components of well-known complexes.

Hubs with Low C

YAL028W	response to stress	molecular_function unknown	endoplasmic reticulum
YBR017C	protein-nucleus import*	nuclear localization sequence binding	cytosol
YBR109C	cytoskeleton organization and biogenesis*	calcium ion binding	cytoplasm*
YBR160W	protein amino acid phosphorylation*	cyclin-dependent protein kinase activity	cytoplasm*
YCR086W	DNA replication*	molecular_function unknown	nucleolus*
YDL017W	protein amino acid phosphorylation*	protein serine/threonine kinase activity	nucleoplasm
YDL030W	nuclear mRNA splicing via spliceosome	RNA binding	snRNP U2
YDL043C	spliceosome assembly	RNA binding	snRNP U2

YDL132W	ubiquitin-dependent protein catabolism*	structural molecule activity*	nuclear ubiquitin ligase complex*
YDL140C	transcription from Pol II promoter	DNA-directed RNA polymerase activity	mitochondrion*
YDR110W	DNA recombination*	ribosomal DNA (rDNA) binding	nucleolus
YDR228C	mRNA polyadenylylation*	protein binding*	mRNA cleavage factor complex
YDR328C	ubiquitin-dependent protein catabolism*	protein binding*	cytoplasm*
YDR395W	mRNA-nucleus export*	protein carrier activity	nucleus
YDR412W	rRNA processing	molecular_function unknown	cytoplasm
YDR477W	protein amino acid phosphorylation*	AMP-activated protein kinase activity	cytoplasm*
YER095W	telomerase-independent telomere maintenance*	recombinase activity	nuclear chromosome*
YER133W	35S primary transcript processing*	protein phosphatase type 1 activity	nucleolus*
YER148W	transcription initiation from Pol II promoter*	DNA binding*	nucleus*
YER165W	regulation of translational initiation	poly(A) binding	cytoplasm*
YER179W	meiosis*	single-stranded DNA binding*	nucleus*
YFL038C	ER to Golgi transport*	GTPase activity	mitochondrion*
YFL039C	cell wall organization and biogenesis*	structural constituent of cytoskeleton*	actin cortical patch*
YGL092W	mRNA-nucleus export*	structural molecule activity	nuclear pore
YGL115W	regulation of transcription from Pol II promoter*	protein kinase activator activity	cytoplasm*
YGL212W	vesicle fusion*	v-SNARE activity	vacuolar membrane (sensu Fungi)
YGL229C	G1/S transition of mitotic cell cycle	protein serine/threonine phosphatase activity	cytoplasm*
YGR074W	nuclear mRNA splicing□ via spliceosome	pre-mRNA splicing factor activity*	small nuclear ribonucleoprotein complex*
YGR172C	vesicle-mediated transport	molecular_function unknown	membrane*
YHR060W	protein complex	unfolded protein	endoplasmic

	assembly*	binding	reticulum membrane
YIL046W	ubiquitin-dependent protein catabolism*	protein binding	nuclear ubiquitin ligase complex*
YIL061C	nuclear mRNA splicing□ via spliceosome	mRNA binding	commitment complex*
YIL144W	chromosome segregation*	structural constituent of cytoskeleton	condensed nuclear chromosome kinetochore*
YIR006C	endocytosis*	protein binding□ bridging	plasma membrane*
YIR009W	nuclear mRNA splicing□ via spliceosome	RNA binding	snRNP U2
YJL030W	mitotic spindle checkpoint	molecular_function unknown	nuclear pore*
YJL203W	nuclear mRNA splicing□ via spliceosome	RNA binding	snRNP U2
YJR022W	nuclear mRNA splicing□ via spliceosome*	pre-mRNA splicing factor activity	nucleus*
YLR116W	nuclear mRNA splicing□ via spliceosome	RNA binding	commitment complex
YLR128W	biological_process unknown	molecular_function unknown	cellular_component unknown
YLR147C	nuclear mRNA splicing□ via spliceosome	pre-mRNA splicing factor activity*	small nuclear ribonucleoprotein complex*
YLR229C	establishment of cell polarity (sensu Fungi)*	GTPase activity*	plasma membrane*
YLR293C	rRNA processing*	GTPase activity	cytoplasm*
YLR368W	mitochondrion organization and biogenesis	molecular_function unknown	mitochondrion
YML064C	signal transduction*	protein binding*	spindle pole body
YMR080C	mRNA catabolism*	ATPase activity*	cytoplasm*
YMR117C	chromosome segregation*	structural constituent of cytoskeleton	condensed nuclear chromosome kinetochore*
YMR138W	microtubule-based process	GTP binding	cytoplasm
YMR308C	mRNA-nucleus export	protein carrier activity	cytoplasm*
YNL189W	nucleocytoplasmic transport	protein carrier activity	cytoplasm*
YNL236W	transcription from Pol II promoter	RNA polymerase II transcription mediator activity	mediator complex
YNL271C	actin filament organization*	cytoskeletal regulatory protein binding	bud neck*

YOL004W	regulation of transcription from Pol II promoter*	histone deacetylase activity	histone deacetylase complex*
YOR036W	Golgi to vacuole transport	t-SNARE activity	Golgi apparatus*
YOR047C	regulation of transcription from Pol II promoter*	protein kinase activator activity	nucleus*
YOR098C	mRNA-nucleus export*	protein binding*	nuclear pore
YOR160W	protein-nucleus import*	nuclear localization sequence binding	cytoplasm*
YOR355W	aerobic respiration	molecular_function unknown	cytoplasm*
YPL031C	protein amino acid phosphorylation*	cyclin-dependent protein kinase activity	nucleus
YPR105C	intra-Golgi transport*	molecular_function unknown	Golgi transport complex
YPR119W	G2/M transition of mitotic cell cycle*	cyclin-dependent protein kinase regulator activity	cytoplasm*
YPR165W	cell wall organization and biogenesis*	GTPase activity*	mitochondrion*

Again, those hubs with low C are predominantly kinases and transporters or otherwise involved in transient binding, as we hypothesized.

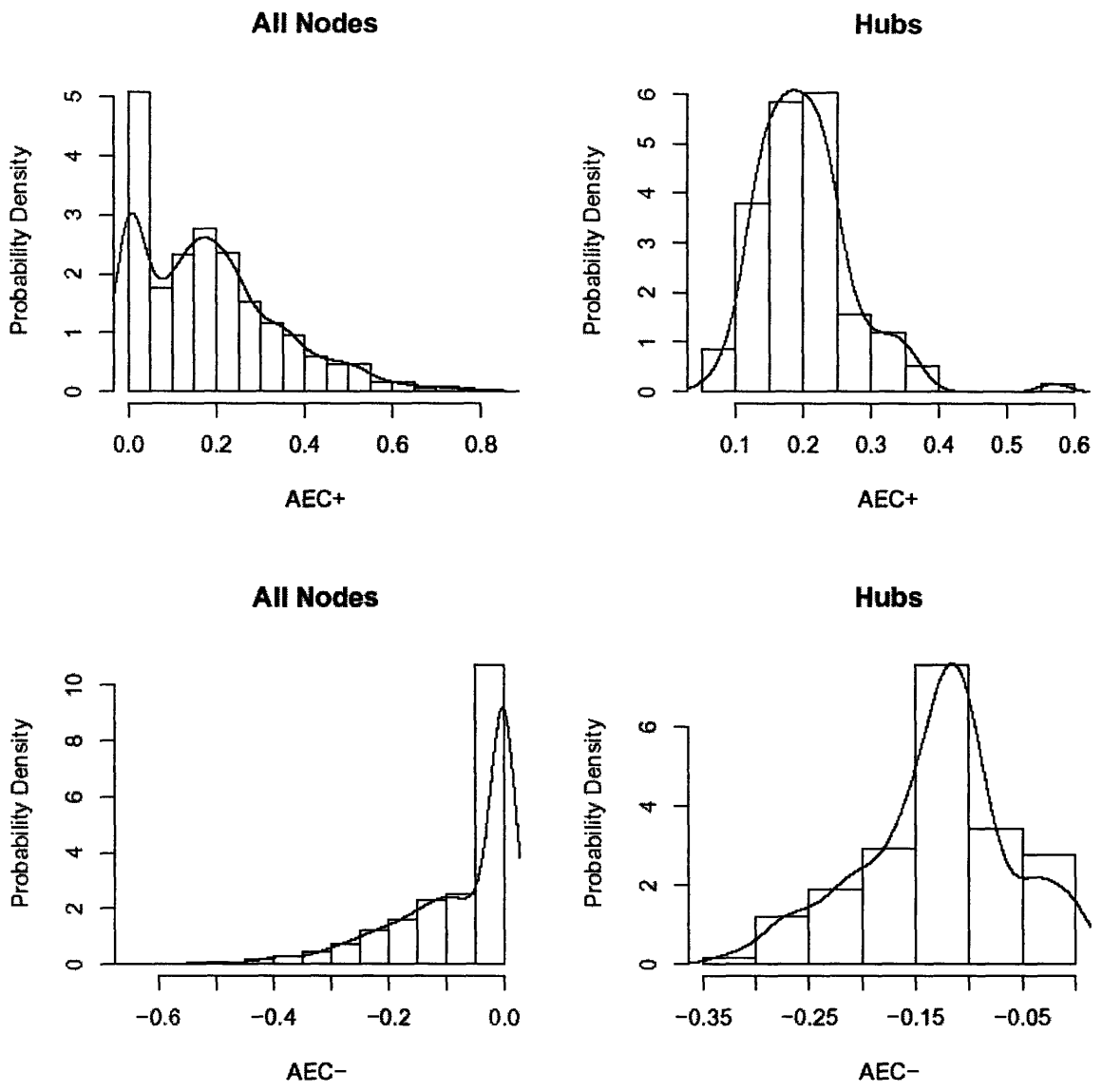
3.8.2 Average Expression Correlation

While all pairwise expression correlations in KSHV are positive, which motivated our definition of AEC, correlations in yeast can be either positive or negative. In particular, around any given node, its neighbors can be either positively or negatively correlated in expression. Thus the AEC defined for KSHV would not directly work for yeast.

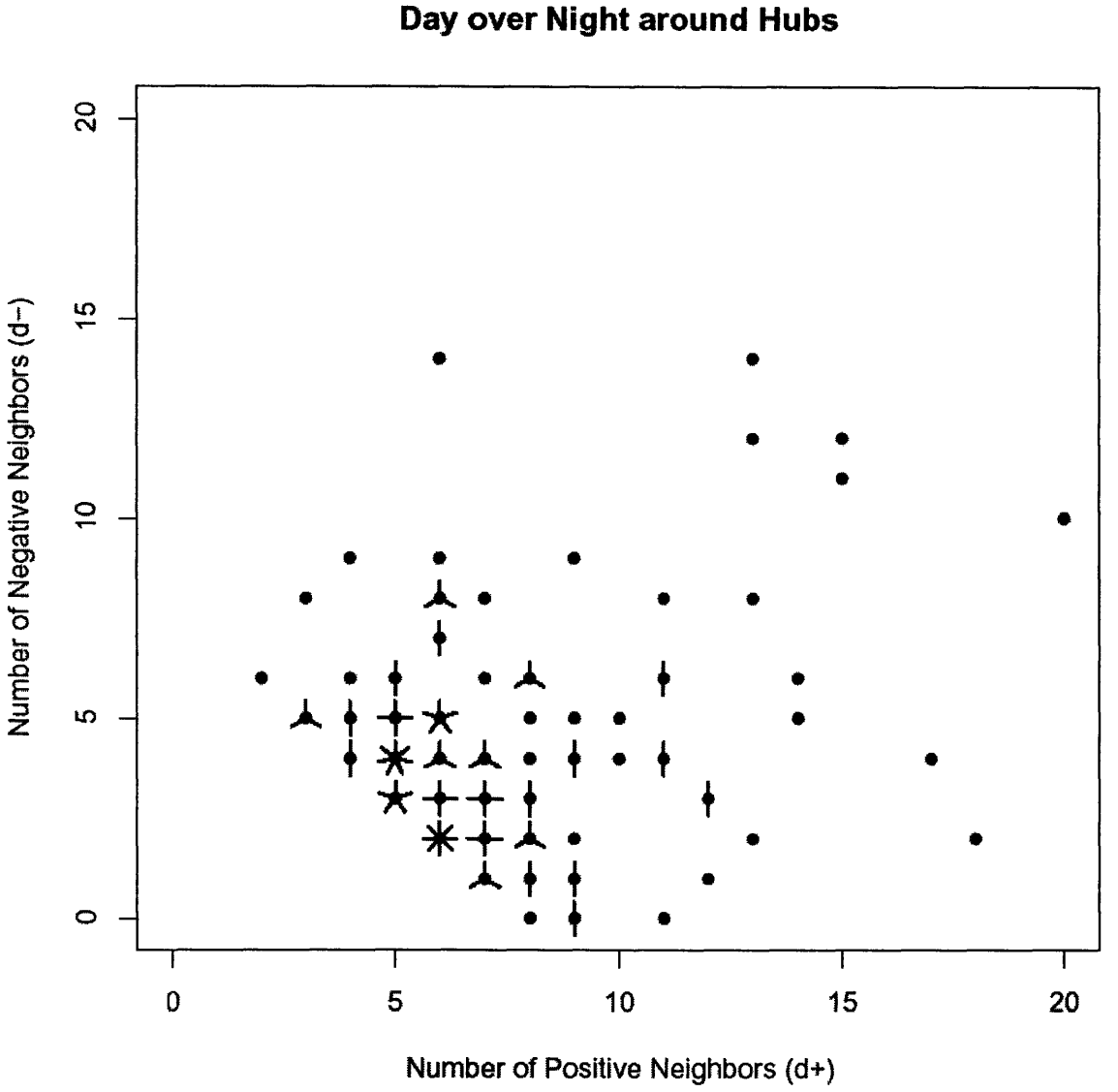
Instead of using a single measure, we look at the positive and negative correlations separately. For a node of degree k, we define k_+ as the number of positively correlated neighbors and k_- as the number of negatively correlated neighbors, with $k = k_+ + k_-$, and define AEC+/AEC- accordingly as the average of positive/negative correlations around that node.

Similar to the analysis with C, now we compare the distributions of AEC+ (respectively AEC-) for all nodes and for hubs only. While the background distribution of all nodes clusters around 0, where noise mostly occurs, the distribution for hubs shifts away from 0 and has a distinct peak on either side. This not only confirms the unique roles played by hubs, but also pinpoints the typical, biologically meaningful values of AEC+/AEC- around the peaks.

Having shown the biological/statistical significance of hubs, we focus the rest analyses on them. First, we note that hubs tend to have more positively correlated neighbors than negatively correlated ones, that is, $k_+ > k_-$ for most hubs ("day over night"), suggesting typical temporal regulation patterns around hubs. While this is not surprising for complexes, those hubs with unusual "day/night" patterns would offer interesting case studies.



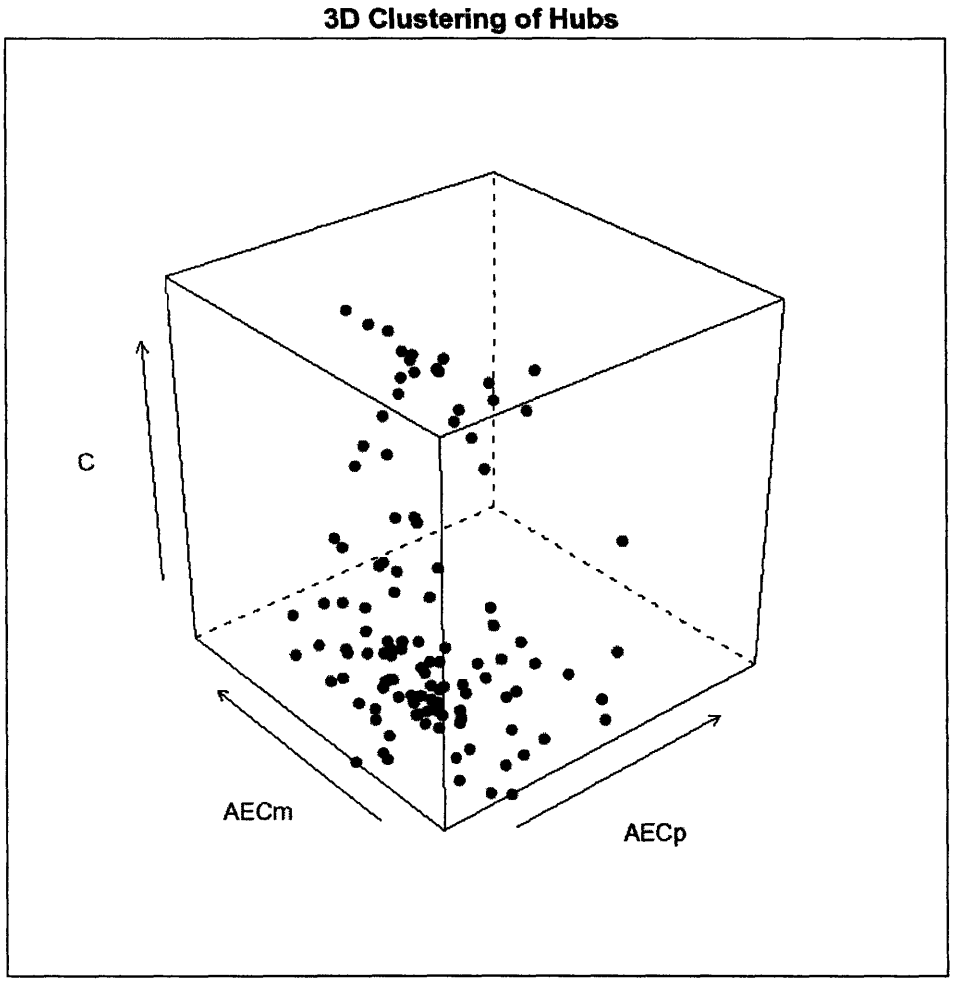
Distribution of AEC+/AEC- for all nodes and for hubs only. Note the distinct peaks for hubs, on both positive and negative sides, after the background noise has been filtered out.



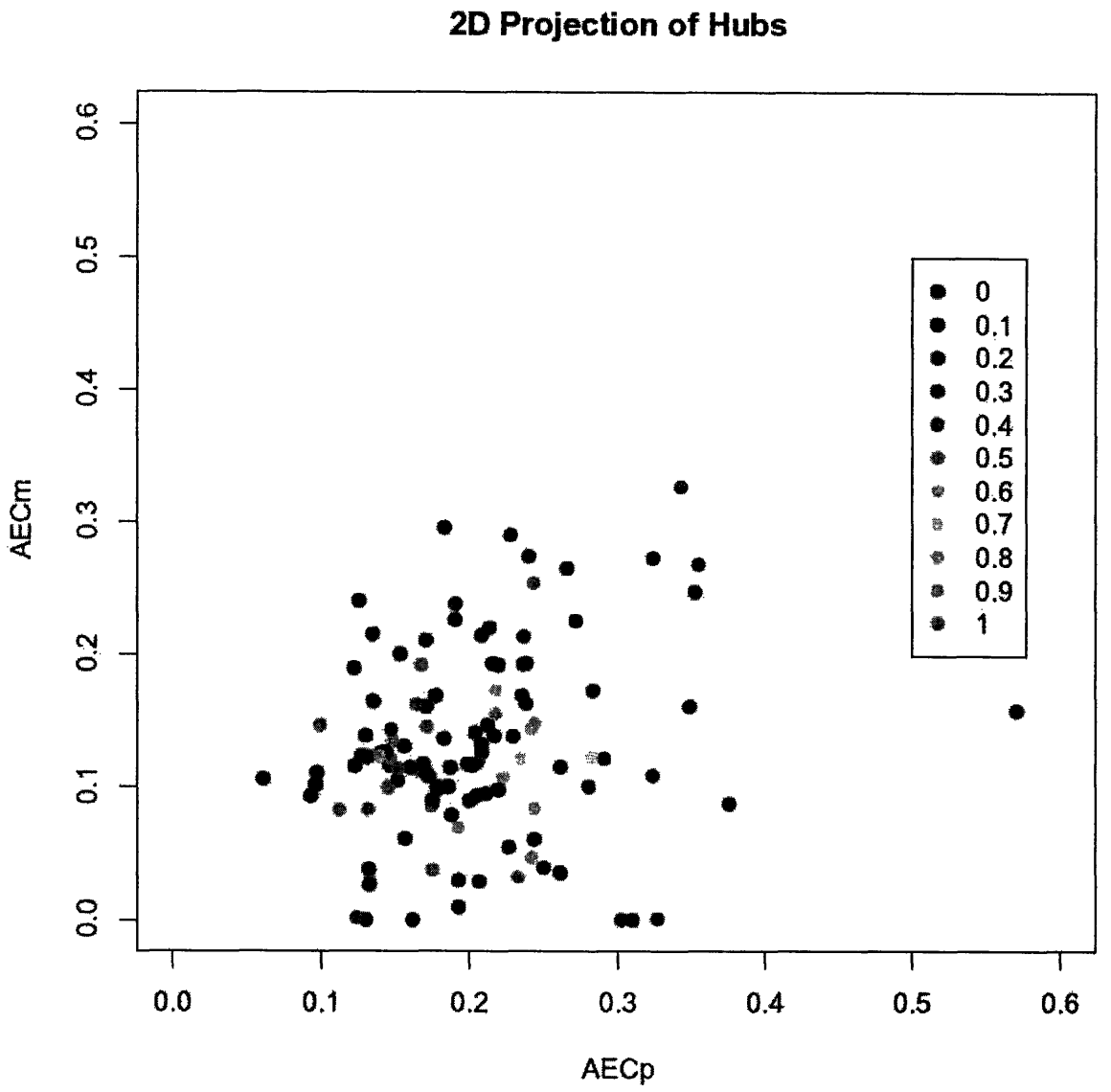
Sunflowerplot of (k_+, k_-) around hubs. The number of observations at each data point, if more than 1, is denoted by the number of stems around the center. As shown in the figure, such points strongly cluster below the diagonal.

3.8.3 Joint Analysis of C and AEC

Having validated C and AEC+/AEC- separately, we now combine them for a joint analysis. We first cluster hubs visually in the C-AEC 3D space and show these two measures are largely orthogonal, then we prove numerically that the combined measure is better at picking out biologically significant correlations than either measure alone.



The 116 hubs are plotted in the C-AEC 3D space. Note that the two measures are not predictive of each other. For a clearer view, we project the points onto the AEC-plane along the C-axis.



For a different view, we project the hubs onto the AEC+/AEC- plane, using color spectrum to denote the corresponding C values. Note how the colors mix without clear boundaries – thus C and AEC are largely independent measures and their combination reveals more than the parts.

To demonstrate this numerically, we compute how often two hubs "close" in space share the same process, function, or component, for different measures of closeness. The background sharing percentage for each biological criterion is computed for all pairwise hubs (6670 of them). Then we define the following three subsets of "close" pairs. We compute the euclidean distance between two hubs:

Close in 3D – distance < 0.1
Close along C – distance < 0.0707
Close along AEC – distance < 0.0707

Thus the distance in 3D is projected onto the C-axis and the AEC-plane, respectively. As the following table shows, distance in 3D performs better than either C or AEC alone, increasing both coverage and accuracy. Interestingly, we also note AEC seems more predictive of process while C more predictive of function and component.

	Random Pair Share	Clustered Pair Share	Fold Enrichment	
Process	0.0299850074963	0.0629750271444	2.10021715527	3D
		0.0370885489105	1.23690310617	C
		0.0442477876106	1.47566371681	AEC
Function	0.0299850074963	0.0499457111835	1.66568946797	3D
		0.0380157626333	1.26782568382	C
		0.0320796460177	1.06985619469	AEC
Component	0.0490254872564	0.0846905537459	1.7274801024	3D
		0.0713954566528	1.45629264793	C
		0.0575221238938	1.17331060053	AEC

Chapter 4 Viral-host Analysis

4.1 Viral-host Interactions in the Literature

While the KSHV viral system is of significant interest on its own, we recognize that viruses do not exist in isolation. Many of their properties, in particular pathogenicity, are meaningful only in the larger context of host interactions. Thus, we would like to connect our KSHV network to a prototypical host network. Unfortunately, herpesviruses mainly infect verterbrates and to date there has been no genome-wide experimental mapping of higher eukaryotic proteomes. Nevertheless, Lehner and Fraser [21] have constructed a first-draft human PPI network, based on orthologous interactions in model organisms. We will use their network as our prototypical host network and investigate the topology of our viral network in that larger context.

To combine the two networks, we need a list of interactions between KSHV and human proteins, that is, the connection edges. Since there is already a vast literature on herpesviruses, we first turned there for interaction data. Aside from helping our own project, a collection of previously scattered herpesviral interactions would be of considerable interest and use to the virology community.

After extracting and reading more than 1000 PubMed abstracts pertaining to herpesviral interactions, we were able to compile a list of ~300 interactions. As expected, the great majority of them are viral-host interactions instead of viral-viral ones. The following table contains those literature interactions between KSHV and human.

PMID	Gene1 (KSHV)	Gene2 (Human)
10200596*	VIRF-2(ORFK11)	ICSBP(ENSG00000140968)
10200596*	VIRF-2(ORFK11)	IRF-1(ENSG00000125347)
10200596*	VIRF-2(ORFK11)	IRF-2(ENSG00000168310)

10200596*	VIRF-2 (ORFK11)	RelA/p65 (ENSG00000173039)
10200596*	VIRF-2 (ORFK11)	p300 (ENSG00000100393)
10377196*	vMIP-I (ORFK6)	CCR8 (ENSG00000179934)
10438822*	vIRF-1 (ORFK9)	ICSBP (ENSG00000140968)
10438822*	vIRF-1 (ORFK9)	IRF1 (ENSG00000125347)
10438822*	vIRF-1 (ORFK9)	p300 (ENSG00000100393)
10438822*	vIRF-1 (ORFK9)	p300 (ENSG00000100393)
10559289*	LANA (ORF73)	RING3 (ENSG00000112526)
10562490*	LANA (ORF73)	HistoneH1 (ENSG00000189060)
10666184*	vMIP-3 (ORFK4.1)	CCR4 (ENSG00000183813)
10736178*	vMIP-II (ORFK4)	CCR5 (ENSG00000188239)
10736178*	vMIP-II (ORFK4)	CXCR4 (ENSG00000121966)
11000236*	LANA-1 (ORF73)	CIR (***)
11000236*	LANA-1 (ORF73)	SAP30 (ENSG00000164105)
11000236*	LANA-1 (ORF73)	mSin3A (ENSG00000169375)
11027294*	vIRF (ORFK9)	p300 (ENSG00000100393)
11038375*	LANA-1 (ORF73)	ATF4/CREB2 (ENSG00000128272)
11090200*	K-bZIP (ORFK8)	p53 (ENSG00000141510)
11160690*	ORF50 (ORF50)	CBP (ENSG00000005339)
11160690*	ORF50 (ORF50)	HDAC-1 (ENSG00000116478)
11160690*	ORF50 (ORF50)	c-Jun (ENSG00000177606)
11336706*	kaposnA (ORFK12)	cytohesin-1 (ENSG00000108669)
11390621*	vIRF1 (ORFK9)	p53 (ENSG00000141510)
11390631*	ORF50 (ORF50)	CBP (ENSG00000005339)
11425857*	LANA (ORF73)	CBP (ENSG00000005339)
11533213*	K8 (ORFK8)	CBP (ENSG00000005339)
11700073*	vMIP-II (ORFK4)	CCR5 (ENSG00000188239)
11711586*	RTA (ORF50)	MGC2663 (ENSG00000130818)
11741976*	RTA (ORF50)	STAT3 (ENSG00000168610)
11752170*	K15 (ORFK15)	HAX-1 (ENSG00000143575)
12388711*	K7 (ORFK7)	CAML (ENSG00000164615)
12477864*	RAP=K8 (ORFK8)	C/EBPalpha (***)
12477864*	RTA (ORF50)	C/EBPalpha (***)
12486118*	LANA (ORF73)	HP1-alpha (ENSG00000094916)
12584338*	gB (ORF8)	alpha3integrin (ENSG00000005884)
12584338*	gB (ORF8)	beta1integrin (ENSG00000150093)
12604819*	K8 (ORFK8)	hSNF5 (ENSG00000099956)
12612078*	RTA (ORF50)	Brg1 (ENSG00000127616)

12612078*	RTA(ORF50)	CBP(ENSG00000005339)
12612078*	RTA(ORF50)	TRAP230(ENSG00000184634)
12768028*	LANA-1(ORF73)	p53(ENSG00000141510)
12829841*	LANA(ORF73)	Gsk-3A(ENSG00000105723)
12829841*	LANA(ORF73)	Gsk-3B(ENSG00000082701)
12832621*	RTA(ORF50)	RBP-Jkappa(ENSG00000168214)
12885907*	Rap=K8(ORFK8)	C/EBPalpha(***)
12885907*	Rap=K8(ORFK8)	p21(ENSG00000124762)
12890756*	vFLIP(ORFK13)	IKK-gamma(ENSG00000073009)
12915577*	K-bZIP(ORFK8)	Cdk2(ENSG00000123374)
12941895???	Lna(ORF73)	KLIP1(***)
9829980*	vBcl-2(ORF16)	DIVA???(ENSG00000137875)

In total there are 53 of them. After filtering out 5 redundancies and 5 interactions where the human interactor does not have an ENSEMBL gene id (denoted by '***'), we are left with 43 viral-host interactions between 14 KSHV and 36 human genes. Among those 36 human genes, 35 do NOT have interactions in the human network! Obviously there isn't enough data to combine the KSHV and human networks.

4.2 Predicting Viral-host Interactions

4.2.1 Motivation

The failure of literature interactions to properly connect the KSHV and human protein interaction networks highlights the urgency for systematic, genome-scale mapping of viral-host interactions, which would be revolutionary in the study of viral pathogenicity. Meanwhile, we propose an innovative approach to predict viral-host interactions *in silico* and use them to connect the KSHV and human networks for a combined viral-host analysis.

While there have been genome-scale interaction mapping for yeast, worm, and fly, interaction data for other organisms remain sparse. Thus we would like to transfer our knowledge of interactions in the three model organisms to a new context, namely the interaction between a virus and its host. The idea is as follows – if a KSHV protein and a human protein both have orthologs in yeast, and those two yeast orthologs interact in yeast, then we consider the pair of KSHV and human proteins a potential viral-host interaction. Similarly, we also map them onto worm and fly networks to see if they interact.

In contrast to previous effort of mapping interactions in one species onto a second one, here we map two different species onto a third one. In order to reliably transfer interactions across, we must have confidence in both the original interactions and the orthology relationships.

4.2.2 Materials and Methods

To ensure the quality of the original source interactions, we constructed a high-confidence, core interaction network for each of the three model organism. For yeast, a core set of interactions were obtained from DIP (Database of Interacting Proteins), as defined by Deane et al.

To identify orthologs between KSHV and yeast, worm, or fly, we used the reciprocal best BLAST hit approach. Consider the long evolutionary distance between herpesviruses and higher eukaryotes, BLOSUM45 were used instead of the default BLOSUM62.

The following table lists the KSHV ORFs with at least one ortholog in the three model organism:

KSHV ORFs	Yeast orthologs	Worm orthologs	Fly orthologs
KSHV_ORF18	***	Y51H4A.17	***

KSHV_ORF2	DYR_YEAST	C36B1.7	CG14887-PA
KSHV_ORF20	***	***	CG7036-PA
KSHV_ORF21	***	F11C3.1	***
KSHV_ORF27	***	***	CG5521-PA
KSHV_ORF36	CHK1_YEAST	***	***
KSHV_ORF39	***	C27D6.8	***
KSHV_ORF4	***	T07H6.5	CG1500-PA
KSHV_ORF46	UNG_YEAST	Y56A3A.29a	***
KSHV_ORF60	RIR2_YEAST	C03C10.3	CG8975-PA
KSHV_ORF61	RIR3_YEAST	T23G5.1	CG5371-PA
KSHV_ORF64	YH17_YEAST	***	***
KSHV_ORF70	TYSY_YEAST	Y110A7A.4	CG3181-PA
KSHV_ORF72	CG24_YEAST	Y38F1A.5	CG9096-PC
KSHV_ORF74	***	***	CG14593-PA
KSHV_ORF75	PUR4_YEAST	F10F2.2	CG9127-PC
KSHV_ORF9	DPOD_YEAST	F10C2.4	CG5949-PA
KSHV_ORFK14	***	ZK377.2b	CG14521-PA
KSHV_ORFK5	SSM4_YEAST	F55A3.1	CG13442-PA

To identify human orthologs in the three model organisms is, however, more complicated. Unlike prokaryotes and micro-organisms, higher eukaryotes have undergone extensive gene duplication events, resulting in multiple potential orthologs in other species. Fortunately, the InParanoid algorithm addresses this issue quite nicely. It first identifies potential orthologs by best pairwise similarity searches, and then clusters those orthologs into groups of likely co-orthologs, with each ortholog assigned a confidence score that it is the main ortholog. We obtained the tables of orthologs between human and the three model organisms from the InParanoid website and extracted only the main orthologs (those with confidence score 1.0) from each orthologous group.

Thus, we have three high-confidence, core interaction networks, and both KSHV and human proteins can be mapped onto the three networks using the six high-confidence tables of orthologs.

KSHV Gene	Human Gene	Swissprot ID	d_{Yeast}	d_{Worm}	d_{Fly}
KSHV_ORF18	ENSG00000008177.4		*	2	*
KSHV_ORF18	ENSG00000034152.4	MPK3_HUMAN	*	2	*
KSHV_ORF18	ENSG00000060688.1		*	2	*
KSHV_ORF18	ENSG00000076554.2	TD52_HUMAN	*	2	*
KSHV_ORF18	ENSG00000094880.1	CC23_HUMAN	*	2	*
KSHV_ORF18	ENSG00000100353.5	IF37_HUMAN	*	2	*
KSHV_ORF18	ENSG00000100522.1		*	2	*
KSHV_ORF18	ENSG00000100632.3	ERH_HUMAN	*	2	*
KSHV_ORF18	ENSG00000104957.2		*	2	*
KSHV_ORF18	ENSG00000109911.5		*	2	*
KSHV_ORF18	ENSG00000111336.3		*	2	*
KSHV_ORF18	ENSG00000111605.6		*	2	*
KSHV_ORF18	ENSG00000111802.2		*	2	*
KSHV_ORF18	ENSG00000112062.4	MK14_HUMAN	*	2	*
KSHV_ORF18	ENSG00000112333.1	NR21_HUMAN	*	2	*
KSHV_ORF18	ENSG00000119242.2		*	2	*
KSHV_ORF18	ENSG00000120341.2		*	2	*
KSHV_ORF18	ENSG00000121022.5		*	1	*
KSHV_ORF18	ENSG00000121083.1		*	2	*
KSHV_ORF18	ENSG00000125676.4	THO2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000126561.4	ST5A_HUMAN	*	0	*
KSHV_ORF18	ENSG00000126945.1	ROH2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000130772.2		*	2	*
KSHV_ORF18	ENSG00000131462.1	TBG1_HUMAN	*	2	*
KSHV_ORF18	ENSG00000134072.1	KCC1_HUMAN	*	2	*
KSHV_ORF18	ENSG00000136827.3		*	2	*
KSHV_ORF18	ENSG00000137693.4		*	2	*
KSHV_ORF18	ENSG00000138382.1		*	2	*
KSHV_ORF18	ENSG00000138443.3		*	2	*
KSHV_ORF18	ENSG00000138663.2		*	2	*
KSHV_ORF18	ENSG00000141552.5	AN11_HUMAN	*	2	*
KSHV_ORF18	ENSG00000143256.1	PFD2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000143314.1		*	2	*
KSHV_ORF18	ENSG00000145675.2	P85A_HUMAN	*	2	*
KSHV_ORF18	ENSG00000148396.3	Y310_HUMAN	*	2	*
KSHV_ORF18	ENSG00000151208.4		*	2	*

KSHV_ORF18	ENSG00000158234.3	FAI1_HUMAN	*	2	*
KSHV_ORF18	ENSG00000160293.4	VAV2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000162378.2		*	2	*
KSHV_ORF18	ENSG00000163106.2	PGD2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000164080.2		*	2	*
KSHV_ORF18	ENSG00000165462.1	PMXA_HUMAN	*	2	*
KSHV_ORF18	ENSG00000165917.2	RAPS_HUMAN	*	2	*
KSHV_ORF18	ENSG00000166902.1		*	2	*
KSHV_ORF18	ENSG00000170312.2	CDC2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000170365.1	SMA1_HUMAN	*	2	*
KSHV_ORF18	ENSG00000172432.4		*	2	*
KSHV_ORF18	ENSG00000173757.2	ST5B_HUMAN	*	0	*
KSHV_ORF18	ENSG00000174444.2	RL4_HUMAN	*	1	*
KSHV_ORF18	ENSG00000176248.1	ANC2_HUMAN	*	2	*
KSHV_ORF18	ENSG00000178127.1	NUHM_HUMAN	*	2	*
KSHV_ORF18	ENSG00000178950.3	GAK_HUMAN	*	2	*
KSHV_ORF18	ENSG00000179912.5	YA02_HUMAN	*	2	*
KSHV_ORF18	ENSG00000182351.3	CRP1_HUMAN	*	2	*
KSHV_ORF18	ENSG00000187391.3	AIP1_HUMAN	*	2	*
KSHV_ORF18	ENSG00000188920.1		*	2	*
KSHV_ORF2	ENSG00000132581.1	SDF2_HUMAN	*	*	1
KSHV_ORF36	ENSG00000005007.2	RNT1_HUMAN	2	*	*
KSHV_ORF36	ENSG00000013275.1	PRS6_HUMAN	2	*	*
KSHV_ORF36	ENSG00000020426.2	MAT1_HUMAN	2	*	*
KSHV_ORF36	ENSG00000051180.3	RA51_HUMAN	2	*	*
KSHV_ORF36	ENSG00000056998.4	GYG2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000065150.3	IMB3_HUMAN	2	*	*
KSHV_ORF36	ENSG00000065427.1		2	*	*
KSHV_ORF36	ENSG00000073536.3	HUS7_HUMAN	2	*	*
KSHV_ORF36	ENSG00000092201.1		2	*	*
KSHV_ORF36	ENSG00000092621.1	SERA_HUMAN	1	*	*
KSHV_ORF36	ENSG00000095002.1	MSH2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000104884.3	XPD_HUMAN	1	*	*
KSHV_ORF36	ENSG00000106355.1	LSM5_HUMAN	2	*	*
KSHV_ORF36	ENSG00000108504.4	CDK3_HUMAN	2	*	*
KSHV_ORF36	ENSG00000110367.2	DDX6_HUMAN	2	*	*
KSHV_ORF36	ENSG00000111987.2	LSM2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000117222.2	RBB5_HUMAN	2	*	*

KSHV_ORF36	ENSG00000117394.4	GTR1_HUMAN	2	*	*
KSHV_ORF36	ENSG00000118520.3	ARGI_HUMAN	2	*	*
KSHV_ORF36	ENSG00000124198.1	BIG2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000130332.4	LSM7_HUMAN	2	*	*
KSHV_ORF36	ENSG00000130520.1	LSM4_HUMAN	2	*	*
KSHV_ORF36	ENSG00000131459.3	GFA2_HUMAN	1	*	*
KSHV_ORF36	ENSG00000132361.3	IF3X_HUMAN	2	*	*
KSHV_ORF36	ENSG00000136813.1		2	*	*
KSHV_ORF36	ENSG00000136936.1	XPA_HUMAN	2	*	*
KSHV_ORF36	ENSG00000145736.3	TFH2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000146092.1	GBLP_HUMAN	2	*	*
KSHV_ORF36	ENSG00000146372.5	HDA2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000149554.2	CHK1_HUMAN	0	*	*
KSHV_ORF36	ENSG00000155229.5		2	*	*
KSHV_ORF36	ENSG00000159352.3	PSD4_HUMAN	2	*	*
KSHV_ORF36	ENSG00000162290.2		2	*	*
KSHV_ORF36	ENSG00000163161.1	XPB_HUMAN	2	*	*
KSHV_ORF36	ENSG00000163754.4	GLYG_HUMAN	2	*	*
KSHV_ORF36	ENSG00000164025.5	ADHX_HUMAN	2	*	*
KSHV_ORF36	ENSG00000164167.1	LSM6_HUMAN	2	*	*
KSHV_ORF36	ENSG00000169100.2	ADT3_HUMAN	2	*	*
KSHV_ORF36	ENSG00000169375.4	SN3A_HUMAN	2	*	*
KSHV_ORF36	ENSG00000170860.1	LSM3_HUMAN	2	*	*
KSHV_ORF36	ENSG00000175324.2	LSM1_HUMAN	1	*	*
KSHV_ORF36	ENSG00000176974.4	GLYC_HUMAN	2	*	*
KSHV_ORF36	ENSG00000183474.2	TFH2_HUMAN	2	*	*
KSHV_ORF36	ENSG00000186298.2	PP1G_HUMAN	2	*	*
KSHV_ORF46	ENSG00000035928.4	RFC1_HUMAN	2	*	*
KSHV_ORF46	ENSG00000049541.2	RFC2_HUMAN	2	*	*
KSHV_ORF46	ENSG00000076242.1	MLH1_HUMAN	2	*	*
KSHV_ORF46	ENSG00000111445.3	RFC5_HUMAN	2	*	*
KSHV_ORF46	ENSG00000113318.2	MSH3_HUMAN	2	*	*
KSHV_ORF46	ENSG00000116062.1	MSH6_HUMAN	2	*	*
KSHV_ORF46	ENSG00000132646.1	PCNA_HUMAN	1	*	*
KSHV_ORF46	ENSG00000133119.2	RFC3_HUMAN	2	*	*
KSHV_ORF46	ENSG00000163918.2	RFC4_HUMAN	2	*	*
KSHV_ORF60	ENSG00000003393.1	ALS2_HUMAN	*	*	2
KSHV_ORF60	ENSG00000100084.4	HIRA_HUMAN	2	*	*

KSHV_ORF60	ENSG00000100242.3	U84B_HUMAN	*	*	1
KSHV_ORF60	ENSG00000105011.1		1	*	*
KSHV_ORF60	ENSG00000109472.1	CBPH_HUMAN	*	2	*
KSHV_ORF60	ENSG00000110042.1		*	*	2
KSHV_ORF60	ENSG00000119715.3	ERR2_HUMAN	*	*	2
KSHV_ORF60	ENSG00000137104.2	GAL7_HUMAN	2	*	*
KSHV_ORF60	ENSG00000138663.2		*	2	*
KSHV_ORF60	ENSG00000139496.4	NUP1_HUMAN	*	2	*
KSHV_ORF60	ENSG00000142507.1	PSB6_HUMAN	2	*	*
KSHV_ORF60	ENSG00000149100.2		*	1	*
KSHV_ORF60	ENSG00000163520.1	FBL2_HUMAN	*	*	2
KSHV_ORF60	ENSG00000166484.5		2	*	*
KSHV_ORF60	ENSG00000168439.5	IEFS_HUMAN	2	*	*
KSHV_ORF60	ENSG00000171848.1	RIR2_HUMAN	0	*	*
KSHV_ORF61	ENSG00000167325.3	RIR1_HUMAN	0	*	*
KSHV_ORF72	ENSG00000004660.4		2	*	*
KSHV_ORF72	ENSG00000056678.5	KFC1_HUMAN	2	*	*
KSHV_ORF72	ENSG00000087586.6	STK6_HUMAN	2	*	*
KSHV_ORF72	ENSG00000094804.1		2	*	*
KSHV_ORF72	ENSG00000100479.1	DPE2_HUMAN	2	*	*
KSHV_ORF72	ENSG00000101558.4		2	*	*
KSHV_ORF72	ENSG00000103044.1	HAS3_HUMAN	2	*	*
KSHV_ORF72	ENSG00000104812.2	GYS1_HUMAN	2	*	*
KSHV_ORF72	ENSG00000105325.3	FZR_HUMAN	2	*	*
KSHV_ORF72	ENSG00000105810.1	CDK6_HUMAN	*	*	1
KSHV_ORF72	ENSG00000108306.2	FL2L_HUMAN	2	*	*
KSHV_ORF72	ENSG00000108504.4	CDK3_HUMAN	1	*	*
KSHV_ORF72	ENSG00000110931.6		2	*	*
KSHV_ORF72	ENSG00000112118.2	MCM3_HUMAN	2	*	*
KSHV_ORF72	ENSG00000113810.4	SMC4_HUMAN	2	*	*
KSHV_ORF72	ENSG00000114978.2		2	*	*
KSHV_ORF72	ENSG00000118689.3	FXO3_HUMAN	2	*	*
KSHV_ORF72	ENSG00000118922.3	KLFC_HUMAN	2	*	*
KSHV_ORF72	ENSG00000119138.1	BTE1_HUMAN	2	*	*
KSHV_ORF72	ENSG00000120438.1	TCPA_HUMAN	2	*	*
KSHV_ORF72	ENSG00000123975.1	CKS2_HUMAN	1	*	*
KSHV_ORF72	ENSG00000134644.3	PUM1_HUMAN	2	*	*
KSHV_ORF72	ENSG00000136933.4		2	*	*

KSHV_ORF72	ENSG00000138346.1	DN2L_HUMAN	2	*	*
KSHV_ORF72	ENSG00000140992.4	PDPK_HUMAN	2	*	*
KSHV_ORF72	ENSG00000151458.2	YB23_HUMAN	2	*	*
KSHV_ORF72	ENSG00000156802.1		2	*	*
KSHV_ORF72	ENSG00000157456.1	CGB2_HUMAN	2	*	*
KSHV_ORF72	ENSG00000163104.5	SRD1_HUMAN	2	*	*
KSHV_ORF72	ENSG00000166851.1	PLK1_HUMAN	2	*	*
KSHV_ORF72	ENSG00000171097.2		2	*	*
KSHV_ORF72	ENSG00000171132.3	KPCE_HUMAN	2	*	*
KSHV_ORF72	ENSG00000175166.3	PSD2_HUMAN	2	*	*
KSHV_ORF72	ENSG00000188186.1		*	*	2
KSHV_ORF72	ENSG00000189285.1		2	*	*
KSHV_ORF75	ENSG00000109736.5		*	*	2
KSHV_ORF9	ENSG00000101868.2	DPOA_HUMAN	2	*	*
KSHV_ORF9	ENSG00000106628.1	DPD2_HUMAN	1	*	*
KSHV_ORF9	ENSG00000132646.1	PCNA_HUMAN	2	*	*
KSHV_ORF9	ENSG00000155636.3		*	*	2

column 1	**kshv gene**
column 2	**human gene (ENSEMBL ID)**
column 3	**human gene (SWISSPROT ID)**
column 4	**distance in yeast network**
column 5	**distance in worm network**
column 6	**distance in fly network**
distance 0	**self-interacting**
distance 1	**directly interacting**
distance 2	**bridged by one other protein**
distance *	**bridged by two or more proteins or not interacting at all**

Note there is little overlap between the predictions from the three model organisms. First of all, this might imply that there are many more viral-host interactions and the analysis done here is far from saturated. Second, this is also a direct consequence of the fact that the yeast, worm, and fly networks themselves have little overlap. This is not surprising, consider the latter two networks are very incomplete. Furthermore, even in the same

organism, the most well studied yeast, there is little overlap between genome-wide datasets produced by different methods (e.g. Y2H versus Mass Spec) or even by the same methods (e.g. Uetz versus Ito, Gavin versus Ho).

4.3 Combined Viral-host Analysis

4.3.1 Motivation

Our KSHV network consists of 115 edges among 50 nodes, and is the first major viral system to date. However, a virus is not an independent, autonomous life form – its crucial features, in particular its pathogenicity, depend on its interaction with its hosts. To put the herpesviral network in perspective, we would like to combine with a host network. Unfortunately, herpesviruses mainly infect vertebrates, and there has been no large-scale protein-protein interaction data for any of the higher eukaryotes to date.

To transfer our current knowledge of interactions in model organisms to other species, Lehner and Fraser [21] have constructed a first-draft human protein-protein interaction network. In their approach, if a pair of human proteins both have orthologs in one of the model organisms and they interact, then the two human nodes are connected by an edge in the human network. To validate their predicted human network, Lehner et al have shown that it preferentially connects proteins that share the same functional annotations.

Due to the importance of understanding herpesviral infection in humans and the availability of the Lehner network, we decided to use the Lehner network as a model host network. To further improve the data, we extracted a core set of high-confidence interactions from that network. The single largest connected component (SLCC) of this core network consists of 10636 edges among 3169 nodes. All subsequent analyses are done using this SLCC and we refer to it as "the human network" from now on.

With a high-confidence viral and a high-confidence host network at hand, we would like to have high-confidence connections between them. Toward this end, the 156 predicted viral-host interactions were filtered, so that only those KSHV-human protein pairs with directly interacting orthologs in one of the model organisms are retained. After the filtering step, we have 20 viral-host interactions between 8 KSHV and 20 human proteins.

KSHV Gene	Human Gene	Swissprot ID
KSHV_ORF18	ENSG00000121022.5	
KSHV_ORF18	ENSG00000126561.4	ST5A_HUMAN
KSHV_ORF18	ENSG00000173757.2	ST5B_HUMAN
KSHV_ORF18	ENSG00000174444.2	RL4_HUMAN
KSHV_ORF2	ENSG00000132581.1	SDF2_HUMAN
KSHV_ORF36	ENSG00000092621.1	SERA_HUMAN
KSHV_ORF36	ENSG00000104884.3	XPD_HUMAN
KSHV_ORF36	ENSG00000131459.3	GFA2_HUMAN
KSHV_ORF36	ENSG00000149554.2	CHK1_HUMAN
KSHV_ORF36	ENSG00000175324.2	LSM1_HUMAN
KSHV_ORF46	ENSG00000132646.1	PCNA_HUMAN
KSHV_ORF60	ENSG00000100242.3	U84B_HUMAN
KSHV_ORF60	ENSG00000105011.1	
KSHV_ORF60	ENSG00000149100.2	
KSHV_ORF60	ENSG00000171848.1	RIR2_HUMAN
KSHV_ORF61	ENSG00000167325.3	RIR1_HUMAN
KSHV_ORF72	ENSG00000105810.1	CDK6_HUMAN
KSHV_ORF72	ENSG00000108504.4	CDK3_HUMAN
KSHV_ORF72	ENSG00000123975.1	CKS2_HUMAN
KSHV_ORF9	ENSG00000106628.1	DPD2_HUMAN

Note that two of the eight KSHV proteins (ORF18 and ORF46) do not have viral-viral PPI in our KSHV network, demonstrating the hidden role of host interaction, which would have important implications on analyses from sequence evolution to network topology.

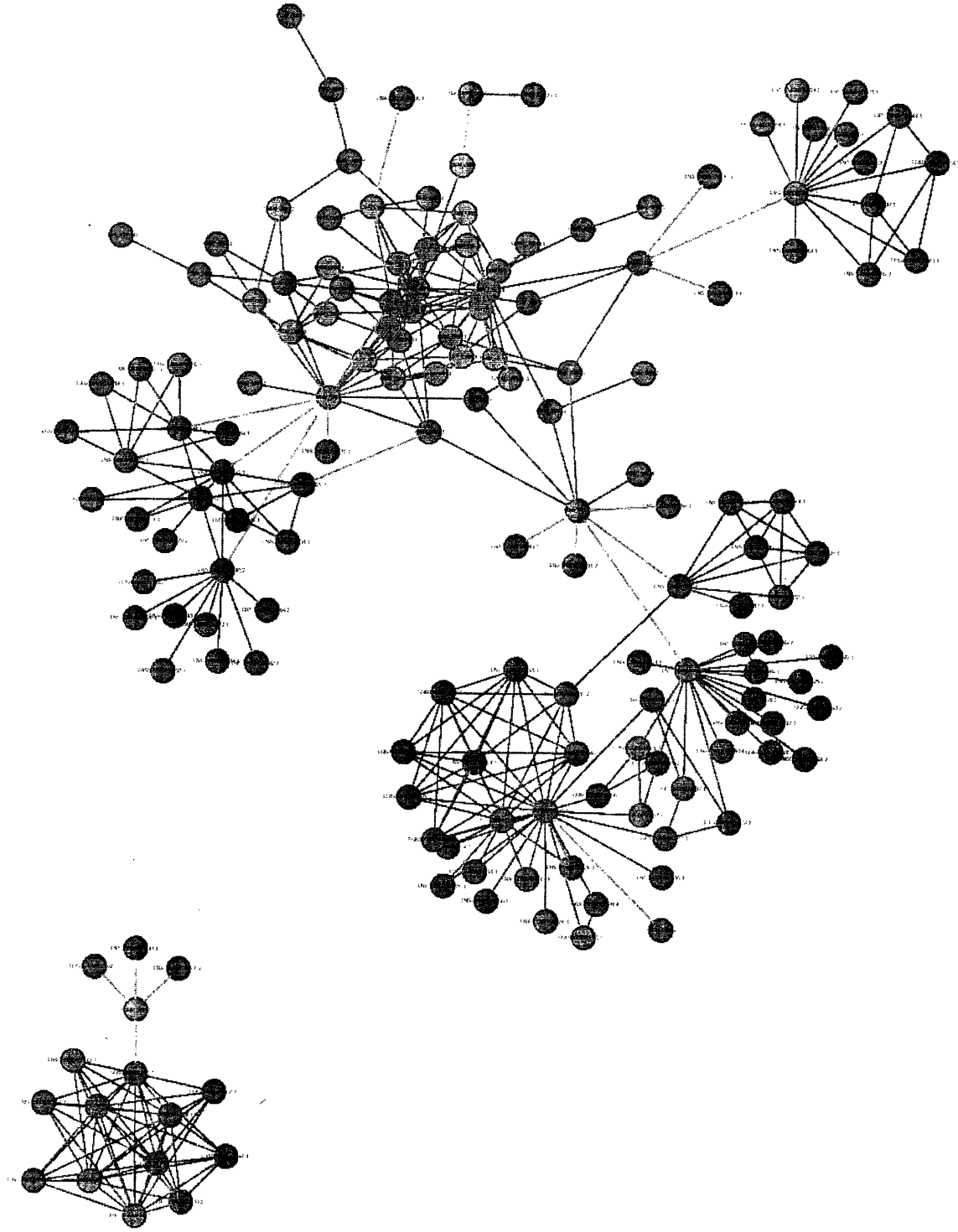
4.3.2 Results

Using the predicted viral-host interactions, we were able to connect the KSHV and human networks into a combined viral-host network. Both the KSHV and the human networks are scale-free, with power coefficient $\gamma = 0.95$ and $\gamma = 1.81$, respectively. The combined network is also scale-free, with $\gamma = 1.82$. As expected, the effect of the human network dominates in the combined one, since the human network is much larger than the viral one (two orders of magnitude).

To isolate the impact of KSHV on human, we zoom out from the KSHV network, one level at a time, into the human network. We define level 1 nodes in the human network to be those human proteins directly targeted by KSHV, and level 2 nodes are level 1 nodes plus their own interaction partners in the human network, and so on. In general, we define level l nodes recursively as level $l-1$ nodes plus all their human interaction partners.

Now we look at the combined viral-host network one level at a time – a viral-host network at level i consists of the KSHV network plus level i human nodes together with their interactions.

Shown is the combined viral-host network at level 2 (that is, KSHV proteins and their human targets plus the human interactors of those human targets). KSHV genes and their interactions are shown in red, human genes and their interactions are shown in blue, while interactions between KSHV and human genes (i.e. viral-host interactions) are shown in green. Note how the topology of the KSHV network changes drastically from a highly coupled module to a more typical scale-free network, where there are distinct modules and crosstalks among them, once the KSHV network is connected to the human network.



4.3.3 Simulations and Discussions

Since herpesviruses attack their hosts and take over cellular machineries to their own advantage, one would expect the combined viral-host network to rapidly take on characteristics of the human network. Thus one measure to assess the quality of the combined network is its scale-free property, in particular the scaling exponent γ.

Aside from asking how much the viral network has improved itself by taking over the host network, we also look at the combined viral-host network from a dual point of view and ask how much damage is done to the human network by the addition of the viral network. One approach is to simply knock out all affected human nodes and the interactions they carry, and look at topological properties of the remaining human network, e.g. the size of its single largest connected component.

Simulations on combined viral-host network

To estimate the statistical significance of network properties of our combined viral-host network, we must construct a suitable null model. The idea behind the construction is that the true, correctly combined network should be able to distinguish itself from random, incorrectly combined networks in terms of network topology.

To construct an equivalent random viral-host network, we generate 20 random "viral-host" interactions. The 8 host-interacting KSHV proteins and their degrees (the number of host proteins they interact with) are both fixed, but their human interaction partners are picked at random. Now we combine the KSHV and the human network using those random connections and analyze the combined viral-host network at each level, in parallell to our analysis on the real viral-host network.

We run 1000 such simulations to generate an ensemble of randomly combined viral-host networks, whose topological properties can then be compared to those of the real network.

The following tables summarize our results, where we compute the size (the number of nodes and the number of edges) and the scaling exponent γ of the single largest connected component of the combined viral-host network at each level, together with their statistical significance (mean, standard deviation, empirical p-value) estimated by the 1000 simulation runs. Similarly, we compute N, E, γ for the SLCC of the rest of the human network, after the viral-affected nodes have been removed, at each level and give their estimated statistical significance.

Level	**Viral-host N**	**Mean**	**SD**	**P-value**
1	65.0	65.08	0.64	0.977
2	146.0	87.62	23.07	0.968
3	506.0	201.54	108.46	0.989
4	1331.0	559.57	315.14	0.994

Level	**Viral-host E**	**Mean**	**SD**	**P-value**
1	133.0	130.1	0.6	0.979
2	288.0	247.1	249.3	0.812
3	1712.0	828.2	765.1	0.841
4	5966.0	2558.8	1717.8	0.979

Level	**Viral-host γ**	**Mean**	**SD**	**P-value**
1	1.1809	1.1749	0.0077	0.970
2	1.5183	1.1364	0.2206	0.999
3	1.3519	1.1271	0.1957	0.913
4	1.4824	1.2548	0.1499	0.966

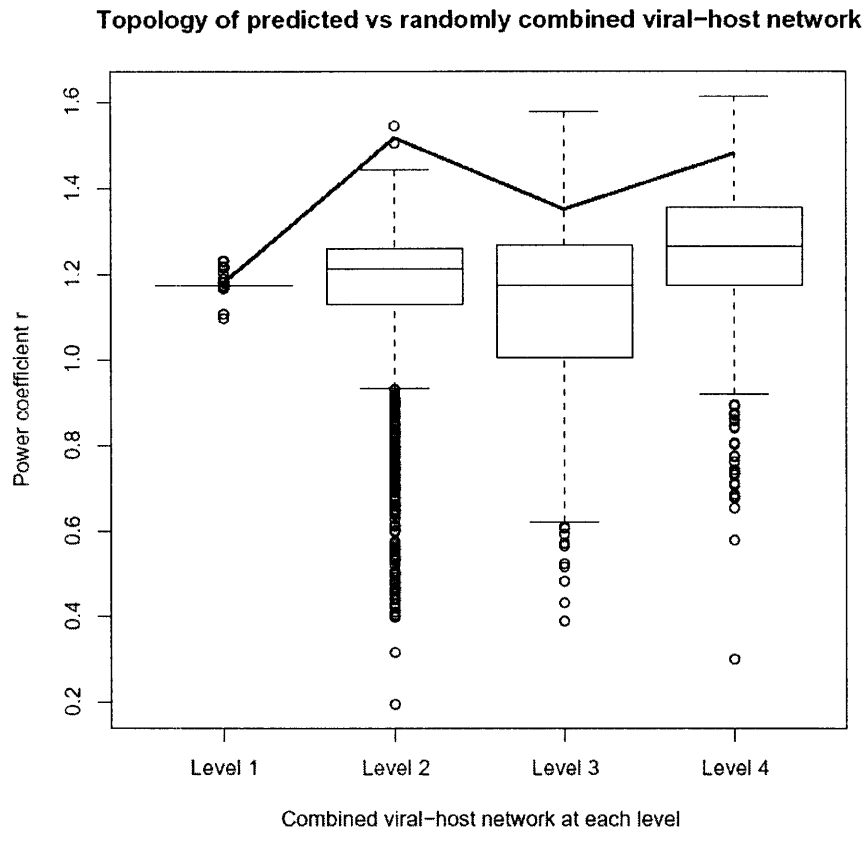
Level	**Host Rest N**	**Mean**	**SD**	**P-value**
1	3145.0	3163.2	4.61	0.009
2	2968.0	3110.0	44.76	0.011
3	2297.0	2899.9	203.71	0.009

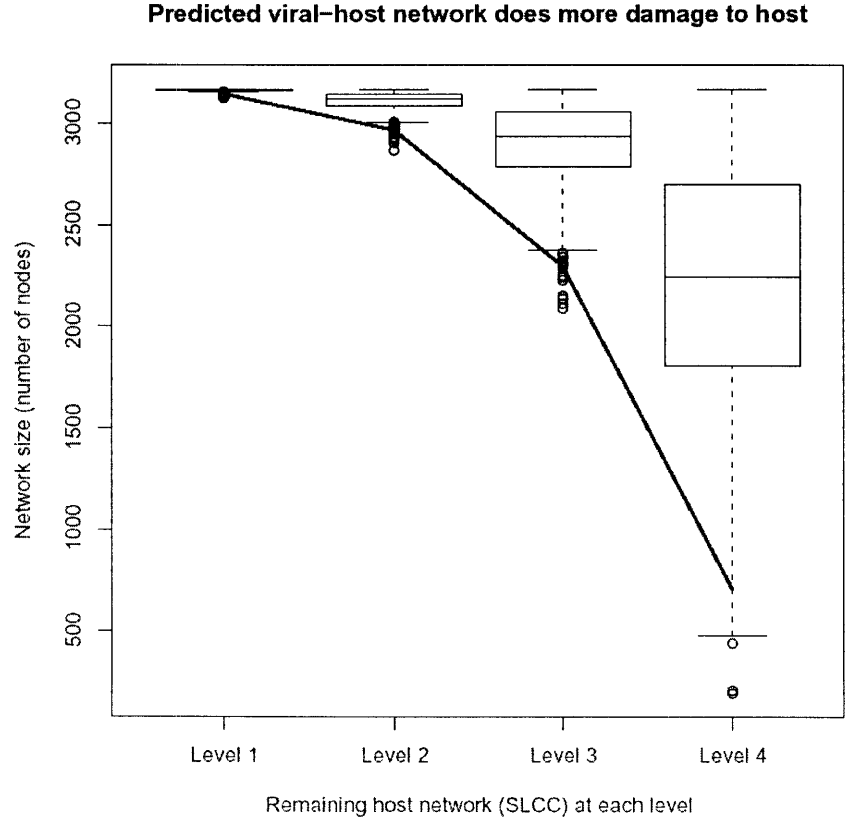
4	708.0	2228.2	604.09	0.010

Level	**Host Rest E**	**Mean**	**SD**	**P-value**
1	10537	10607	24.7	0.023
2	9852	10183	548.6	0.164
3	5787	8998	1320.8	0.018
4	1513	6132	2514.9	0.014

Level	**Host Rest γ**	**Mean**	**SD**	**P-value**
1	1.8264	1.8080	0.0093	0.949
2	1.7811	1.8232	0.0483	0.042
3	2.0274	1.8447	0.0771	0.966
4	1.8673	1.8753	0.1066	0.503

To put the numbers in perspective, for each of the above six topological parameters, we visualize the corresponding result as follows: The distribution of random parameters at each level is shown as a boxplot (the middle line in the box is the median, the box itself corresponds to the spread, while whiskers and points further out are outliers). The four boxplots, corresponding to the four levels, are shown side by side, with the polygon line connecting the true parameter at each level. Two representative parameters and the corresponding figures are shown below:


Topology of predicted vs randomly combined viral−host network
Power coefficient r
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Level 1
Level 2
Level 3
Level 4
Combined viral−host network at each level


Predicted viral−host network does more damage to host
Network size (number of nodes)
500
1000
1500
2000
2500
3000
Level 1
Level 2
Level 3
Level 4
Remaining host network (SLCC) at each level

Note the dramatic difference just by how we combine the two networks – all difference results from the identity of KSHV targets in human. In comparison to randomly combined networks, our viral-host network is significantly bigger in size at each level. Furthermore, at each level our network has a bigger power coefficient, that is, is more similar to the human network in terms of scale-free topology.

As the level goes up, the combined network should have an increasingly bigger power coefficient, as it takes on more and more characteristics of the much larger human network. We indeed observe such a trend for our own viral-host network (except for the spike at level 2), while the power coefficients for random networks remain random and flat as the level goes up.

Intuitively one would expect the combined network at level 2 to have the highest impact, since KSHV proteins affect not only those human proteins they directly interact with, but also their interaction partners down the chain. As the level goes further up, many more human proteins are drawn in and the effect becomes diluted and less specific. The sharp spike at level 2 for our own viral-host network ($p = 0.001$ compared to random ones) supports this view and is thus actually quite reassuring.

Taken together, the dramatic difference between the predicted viral-host network and those randomly combined ones not only validates the predicted viral-host interactions as being likely correct and their human targets as being special, but also shows that network topology is indeed a key determinant of viral-host interactions and viral pathogenicity.

4.3.4 Further Analysis

Simulations under more stringent conditions

While in the above simulations the human targets are chosen uniformly at random from human genes with orthologs in at least one of the three model organisms, the question remains whether the orthologous mapping procedure we employed in predicting viral-host interactions has hidden bias towards, for example, selecting more human targets from within the human PPI network or selecting human targets with higher connectivity ("hub effect"). Now we address this question on two different levels – first we consider each intermediate model network separately and show that there is minimal hidden bias, then we run simulations from ground up by repeating the whole orthologous mapping procedure on the superimposed network and show that it is the identity of KSHV orthologs (predicted versus randomly assigned) that is responsible for the distinct topology of the combined viral-host network.

First, the human network (SLCC) consists of 3169 nodes, with an average degree of 6.7, while there are ~7500 human genes with orthologs in at least one of the three model organisms. Thus, in the previous simulations where the 20 human targets are chosen uniformly at random, ~40% of them would fall within the human network, with an average degree of 6.7 – this compares to the 20 predicted human targets, where 11 (55%) lie within the human network, with an average degree of 10.2.

Since the predicted human targets are mapped from one of the model networks, we look at each of these networks separately to see if the mapping procedure introduces any hidden bias.

The yeast network consists of 2624 nodes, among which 1222 have orthologs in human (i.e. mappable). However, among those 1222 nodes, only 406 (33.2%) can be mapped onto the human network, while the rest fall outside. The 406 mapped human targets have an average node degree of 10.4 in the human network. Thus, while the mapping procedure using yeast as an intermediate does select for hubs, it cannot explain the number of human targets selected ($p < 0.05$ under binomial distribution).

We repeat the same analysis for the worm and fly networks and have the following results:

	Network size	Mappable	Within	% within	Avg deg
Yeast	2624	1222	406	33.2%	10.4
Worm	1415	530	217	40.9%	6.8
Fly	4190	1612	529	32.8%	6.8

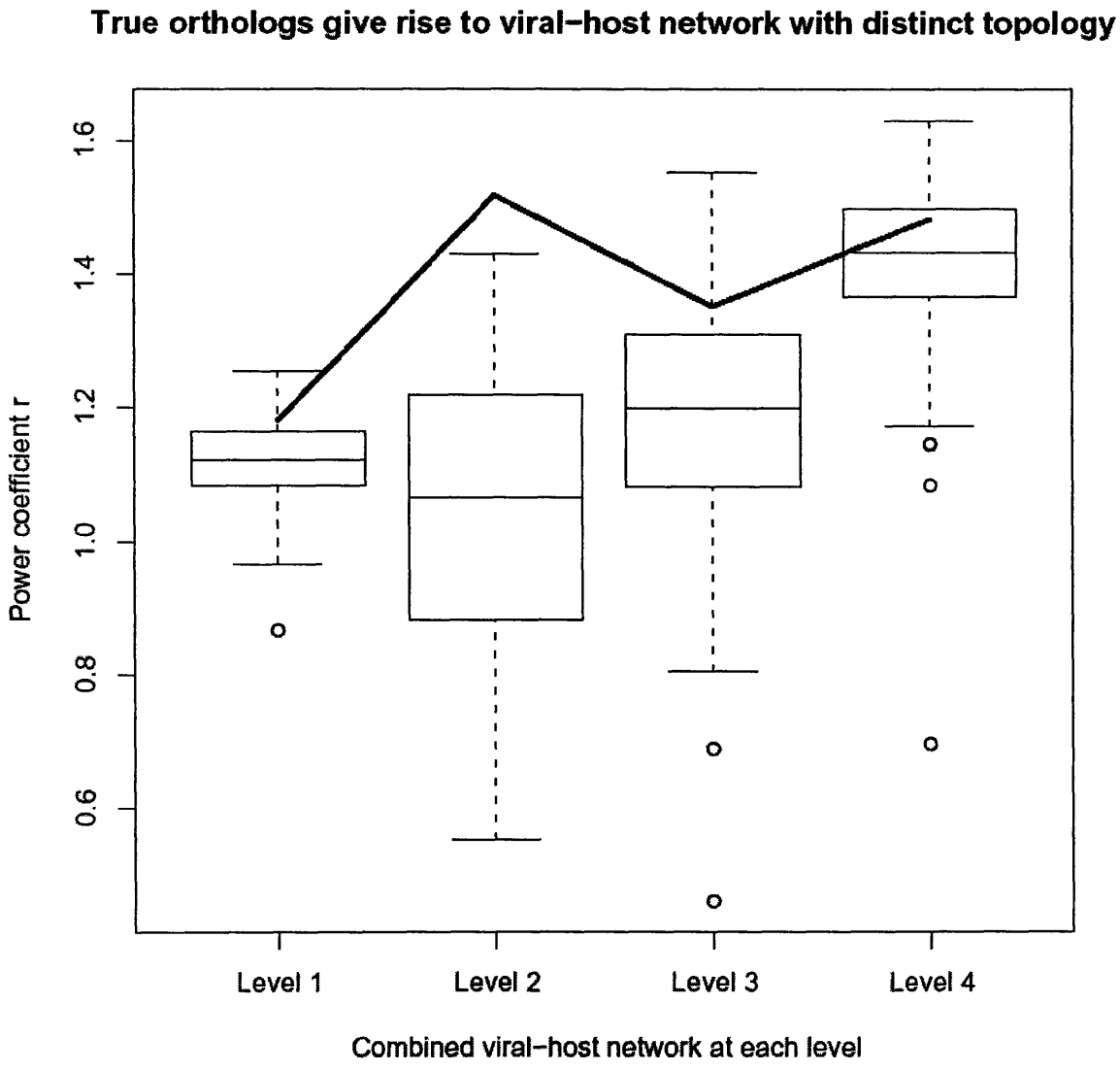
Thus, mapping through the worm or fly network selects for neither the number nor the connectivity of human targets, while mapping through the yeast network enriches the connectivity but not the number of human targets.

Simulations from ground up

Since the human network itself is an orthologous superposition of the three model networks and mapping through the three intermediates simultaneously could introduce hidden bias in complicated ways, we run simulations from ground up to validate the predicted KSHV orthologs, KSHV-human interactions, and combined viral-host network.

Among the KSHV ORFs, 11 have predicted orthologs in yeast, 6 of which lie within the yeast network. Similarly, 14 KSHV ORFs have predicted orthologs in worm, 2 of which lie within the worm network; the corresponding numbers for fly are 13 and 5. We fix those 6, 2, and 5 KSHV ORFs and assign to them at random "orthologs" in the yeast, worm, and fly networks, respectively. Note those three sets of KSHV ORFs need not be disjoint – if a KSHV ORF has an ortholog in each of the three model networks, then we assign to it a random "ortholog" in each network.

After this random assignment of "orthologs", we repeat the whole orthologous mapping procedure as in the prediction of viral-host interactions: If any "ortholog" in any of the three model networks has a neighbor with an ortholog in human (not necessarily within the human network), then we "predict" an interaction between the corresponding KSHV and human genes. The results from all three model networks are merged to give a unique set of "predicted" viral-host interactions.



Thus, even under the most stringent simulation criteria, where the identity and interaction patterns of human-targeting KSHV ORFs are fixed and only their "orthologs" are assigned at random, the key conclusion from our previous analysis continues to hold: The predicted viral-host network at level 2, the biologically meaningful level, is significantly different from simulated ones in the key parameter of scale-free networks.

Chapter 5 Large-scale Retest of Y2H Interactions

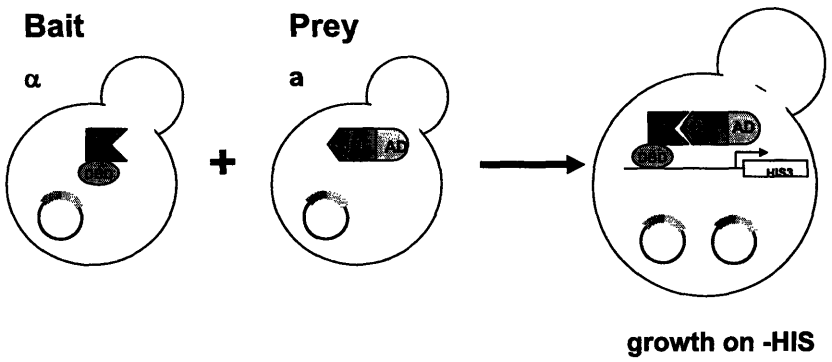
5.1 Introduction

Proteins and their interactions form the molecular basis of life. They form structural or functional complexes, which are the main working machineries in a cell, and they constitute signal transduction pathways, passing along information crucial to the cell's survival, from both inside the nucleus and outside the cellular membrane. Thus, in order to understand the working of a cell and life itself, it is of critical importance to detect and understand protein-protein interactions (PPI).

Currently there are several experimental methods to detect PPI, with relative strengths and weaknesses. Structural approaches like X-ray crystallography or NMR offer the best resolution – not only are interactions unambiguously confirmed, but the interaction interface is also available, from which key residues determining the interaction specificity can often be detected and experimentally confirmed by point mutational analysis. However, solving for 3D structure is an expensive, time-consuming, and sometimes technically infeasible approach. While the data in GenBank, i.e. genomic sequences, have been on an exponential growth course in the last 10 years, the structural data in Protein Data Bank remain modest. Despite current initiatives in structural genomics, solving for 3D structures will not become a routine procedure to detect PPI in the foreseeable future.

Current methods to detect PPI on a genomic scale are either proteomic or genetic in nature. The former includes CoIP/MassSpec and protein chips, while the most famous example of the latter is yeast two-hybrid (Y2H).

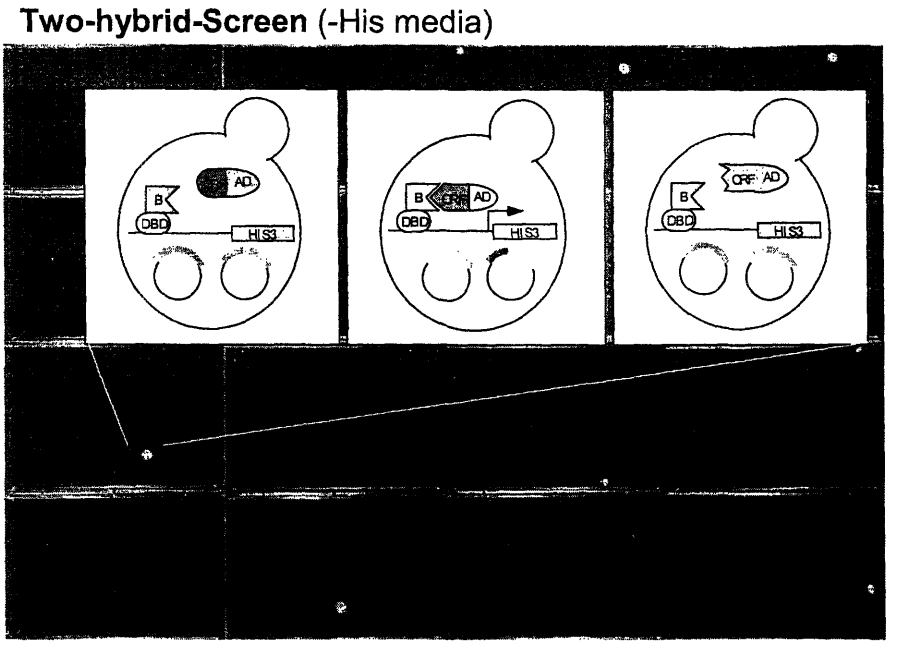
Two-hybrid Principle:



The Y2H system is a split-transcription factor system. The idea is as follows: To test if two ORFs A and B interact, we fuse A with a DNA-binding domain (DBD) and B with an activation domain (AD) and express them in yeast haploids of the opposite mating type. After mating, if A and B indeed interact, they would bring together the DBD and AD and reconstitute the original transcription factor, which would then turn on a reporter gene and enable the diploid to grow on selective medium. Since ORF A (the one fused with DBD) sits on the promoter region, waiting to be activated by a certain ORF B (the one fused with AD), A is called the bait and B is called the prey.

The principle as illustrated only tests a single pair of ORFs for interaction. To enable interaction mapping on a genomic scale, Ito et al have developed the pool approach, while Uetz et al have pioneered the array approach. In the pool approach, we mate a pool of x baits with a pool of y preys and then select for positives and sequence the inserts to obtain the identity of the baits and preys. Thus, all xy pairwise interactions are tested, but the identity of positives is not known beforehand. In contrast, we mate a single bait

against an array of preys (e.g. all ~6000 yeast ORFs) in the array approach, and the identity of positives can be read out directly from their position on the array.



Array Y2H: one bait against all 6000 preys Uetz et al. Nature 2000

In the illustration, one bait (same B) is mated against all 6000 preys (represented by ORFs of different shapes). While the great majority of pairs do not interact, some pairs do (e.g. the middle panel) and will show up positive on the array (the white dot).

Despite its efficiency and success, Y2H does have some serious limitations, including variability due to differential expression of reporter genes, false positives due to self-activating baits or random noise, and false negatives on technical ground (e.g. hybrid fusion proteins fail to fold or localize properly, interaction surfaces are blocked by fused domains, extra cellular signals are needed). Indeed, some of these issues, especially the false positive issue, have presented major problems. It has been shown that large-scale Y2H datasets can contain up to 90% false positives, while even the better ones have about 50% false positives. Since such large-scale datasets constitute the major source of PPI in public databases and all subsequent analyses, either experimental or bioinformatical, are done on and conclusions drawn from such data, it is of crucial

importance to filter out as many false positives as possible from the very beginning, before they propagate further downstream.

In this study, we investigate the array method of Y2H and show that reproducibility is the key to filtering out false positives. Along the way, we confirm a set of high-confidence, previously unknown interactions and explore their biological significance. Furthermore, we estimate the false positive rate for our own data and compare it to other large-scale datasets.

5.2 Materials and Methods

Over the past years, we have accumulated over 1500 reproducible (double positive in the same screen), specific (positive with non-promiscuous prey) Y2H interactions through several hundred independent screens. Conscious of the large number of false positives, we decided to carry out a second, independent screen to retest all those putative positives, before we make them available to the large biology community.

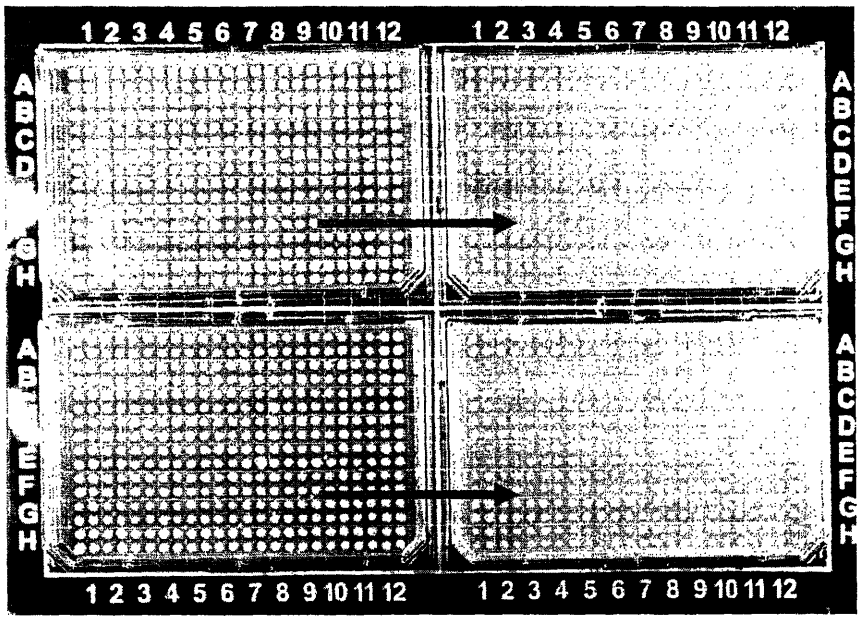
After filtering out baits or preys unavailable on technical ground, we are left with 998 interactions among 272 baits and 706 preys. To find out what is already known in public domain, we looked the Database of Interacting Proteins (DIP), one of the most comprehensive sources of PPI data. The yeast interactions in DIP fall into two classes – CORE and Y2H, the former are confirmed by small-scale or multiple experiments and thus more confident, while the latter come from large-scale two-hybrid screens and are confirmed only once. We looked at the intersection of our dataset with DIP:

Class	Number of Interactions
Novel	380
CORE	132
Reverse CORE	112
Y2H	232
Reverse Y2H	142
Total	998

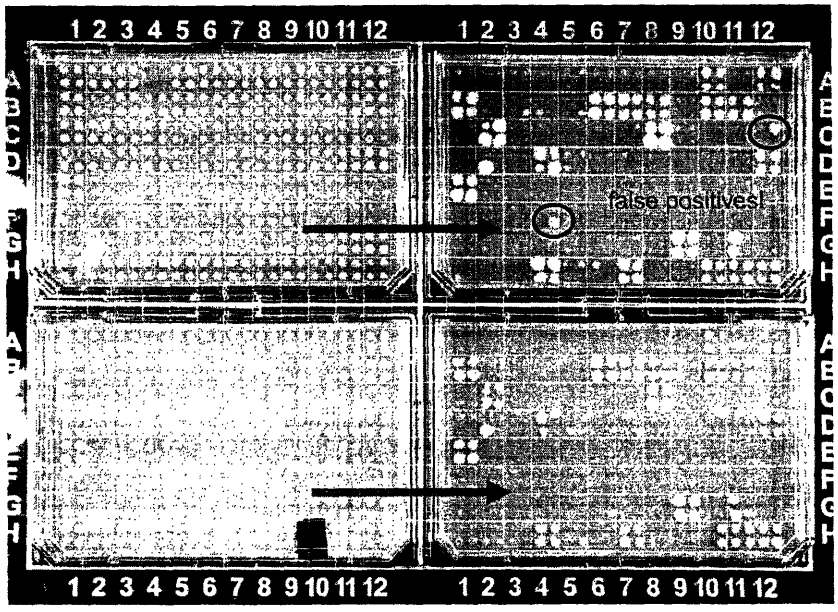
Due to the well-known asymmetry of Y2H, that is, two ORFs are shown to interact in the bait-prey order but not the other or vice verse, we treat A-B and B-A as two different pairs of interactions. Thus, for a pair A-B from our own dataset, either it is not in DIP at all (novel), or it is in either CORE or Y2H, or its reverse is in either CORE or Y2H. Thus, the novel and Y2H classes of interactions are of primary interest, while the CORE class provides positive control, and the two reverse classes will be used to investigate the asymmetry issue. To further ensure the quality of our experimental results, we also use bait-specific negative controls – there are ~1300 preys that have never shown up positive with any bait screened so far – we use them as negative controls; the more partners a bait has, the more negative controls we use. Furthermore, we test all interactions in duplicates of four (instead of the routine), since our experiment is as much about validating methods as discovering new interactions.

To fit our agenda within the framework of traditional array screens is, however, not a trivial issue. Considerable computational and experimental resources were devoted to the experimental design of this project, which falls outside the scope of this thesis and we omit here. Instead, we present some typical array plates demonstrating some of the key issues of the array Y2H approach: successful mating, false positives, activators, and bona fide interactions.

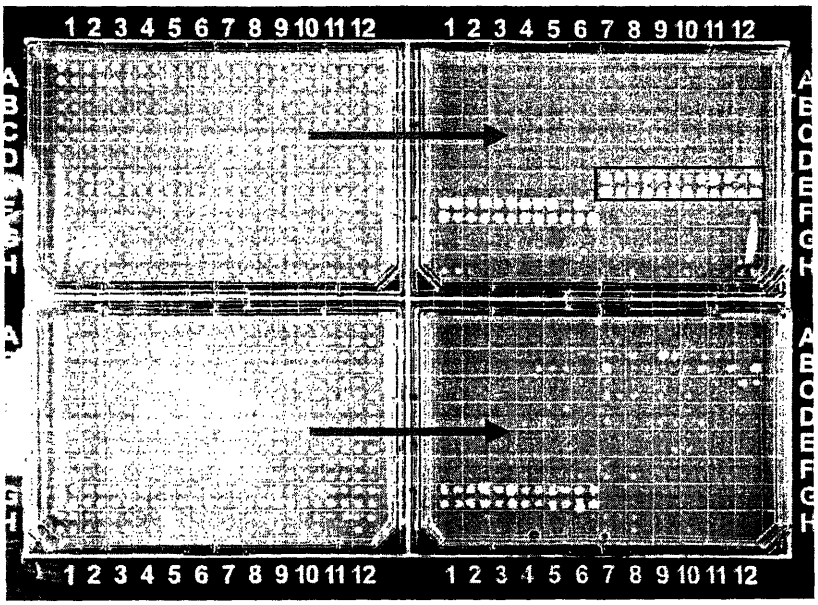
Snapshots along the way



Matings on YEPD Matings on -LW



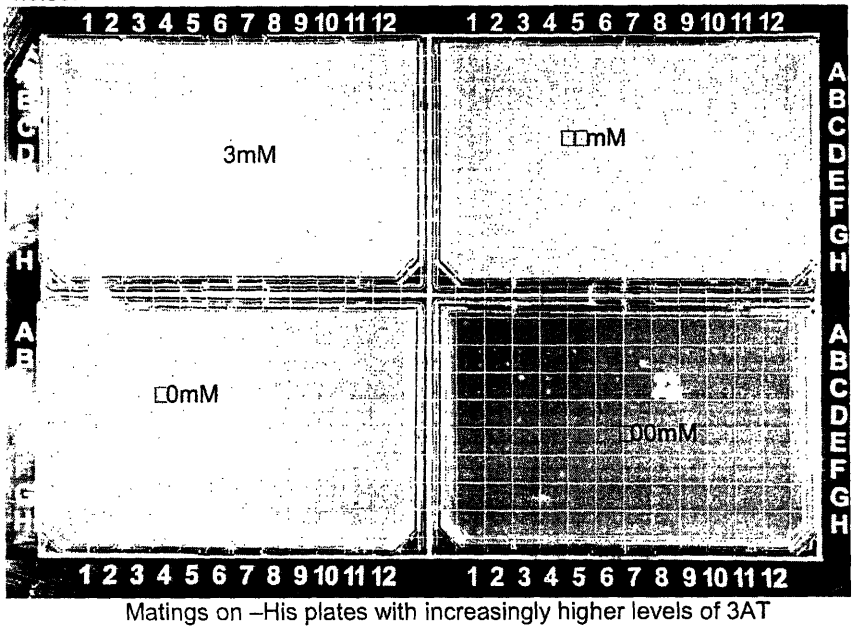
Matings on –LW Matings on –His



Matings on –LW Matings on –His

Negative control plates clearly show which baits are activators!

For activators, use higher levels of 3AT to differentiate true interactions from noise.



Matings on –His plates with increasingly higher levels of 3AT

5.3 Biological Results and Discussions

Class/Size	All	Matings OK	Retest Positive	Percent
Novel	380	306	27	9
CORE	132	102	27	26
Reverse CORE	112	86	24	28
Y2H	232	196	16	8
Reverse Y2H	142	132	16	12
Total	998	822	110	13

Thus, we are able to confirm a much higher fraction of high-confidence, known interactions than novel ones, and there is indeed a difference between Y2H and Reverse Y2H.

Experimentally, we are able to report 27 high-confidence, previously unknown interactions and confirm 32 previously unreliable Y2H interactions. The new interactions are summarized in the following table:

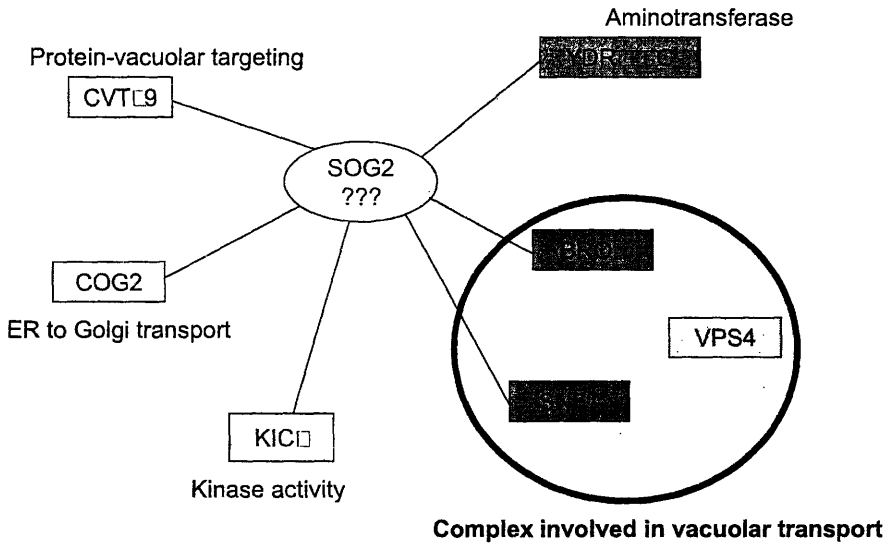
Bait	**Bait function/role**	**Prey**	**Prey function/role**
THP2	Nucleic acid binding; DNA recombination; RNA elongation (Pol II); mRNA-nucleus export	YGR179C	Chromosome segregation
THP2		NUT2	Pol II transcription
THP2		MED7	Pol II transcription
THP2		VPS17	Vesicular transport
SOG2	Unknown	SNF7	Vacuolar transport
SOG2		YDR111C	Amino-acid metabolism
SOG2		BRO1	Small molecule transport
MDM30	Unknown	DNA43	DNA systhesis
MDM30		SPS18	Meiosis
MDM30		SGT1	Protein degradation
SGS1	Chromosome segregation	YLR415C	Unknown
SGS1		DUN1	Mitosis
YAL028W	Unknown	NBP1	Chromatin structure
YAL028W		AKR1	Signal transduction
YPT6	Vesicular transport	YPL192C	Karyogamy
YJL097W	Unknown	YPL192C	
YNL146W	Unknown	YPL192C	
FRQ1	Calcium binding	YSP3	Protein degradation

MPC54	Spore wall assembly	MPC54	Spore wall assembly
HUB1	Protein tagging	SNU66	RNA splicing
KRI1	Ribosome biogenesis	RPN8	Protein degradation
SEC17	ER to Golgi transport	YOL010W	Unknown
ECO1	DNA repair	MPS3	Nuclear migration
YLR128W	Unknown	YNL247W	Protein synthesis
YOL022C	Unknown	ROD1	Cell stress
YJL048C	Unknown	YJL048C	Unknown
YMR269W	Unknown	YMR269W	Unknown

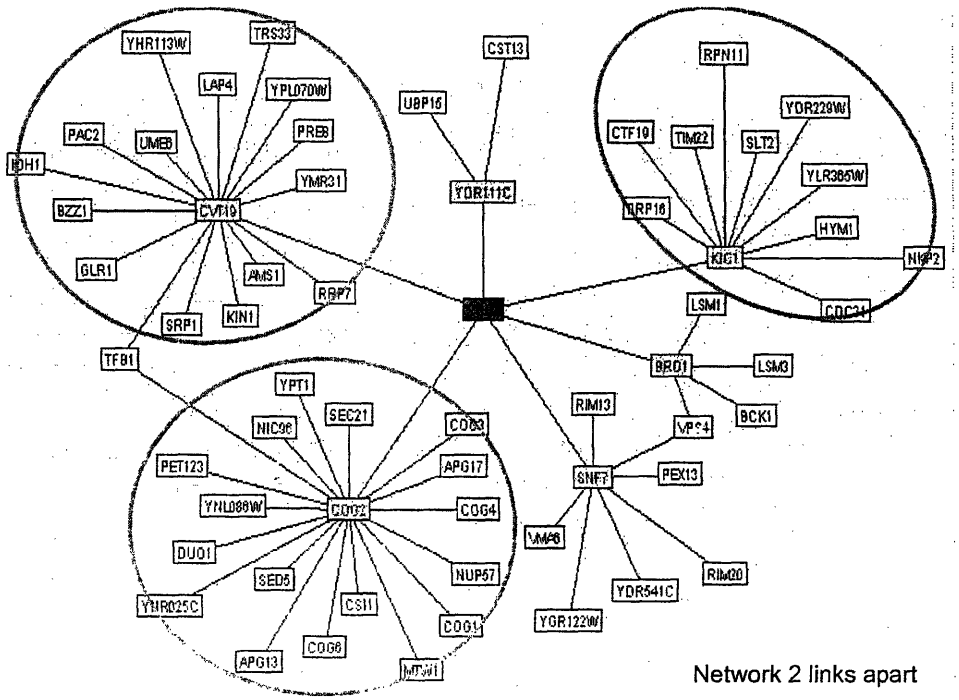
Thus, even for yeast, the most well-studied model organism, much remains unknown!

Now we proceed and discuss some of the biological significance of the interactions we discovered. For example, SOG2 is an essential gene, yet to date nothing is known about it – none of its biochemical function, localization, or cellular role is known. Previously it has been shown to interact with CVT19, COG2, and KIC1, three proteins of diverse roles, which provides little clue to the role of SOG2 itself. Though our retest, we have discovered that it also interacts with YDR111C, BRO1, and SNF7, with high-confidence. Furthermore, it is known that the latter two, together with VPS4, form a complex involved in vacuolar transport. Thus our analysis has associated SOG2 with the vacuolar transport process, with its action possibly modulated by the two enzymes.

SOG2: essential, yet absolutely nothing is known!



To further illucidate the context of protein-protein interactions around SOG2, we zoom out one more level and look at its neighbors's neighbors.

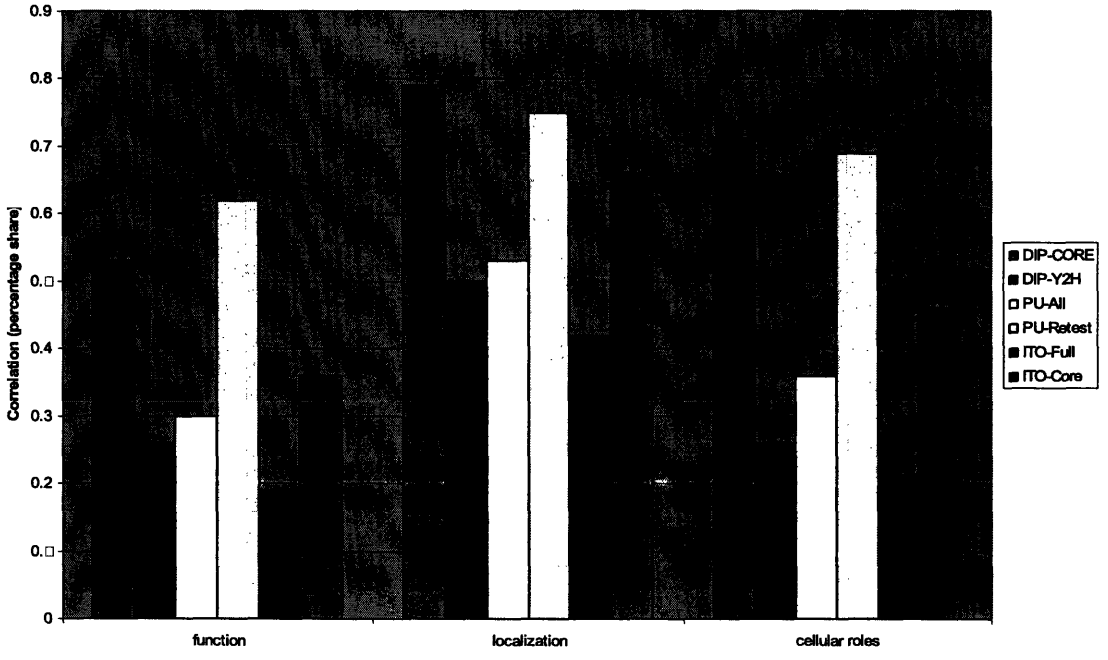


5.4 Statistical Results and Discussions

Now we evaluate the quality of our dataset using several independent statistical measures and compare it with other large-scale datasets. In particular, we derive the false positive rate of them.

Mrowka et al have used mRNA expression correlation to evaluate the quality of genome-scale PPI datasets. Here we introduce three other measures, namely the fraction of pairs in the dataset that share function, localization, or cellular role, and use them to analyze the following datasets:

DIP-CORE, DIP-Y2H	Most comprehensive compilation
PU-All, PU-Retest	Our own data, before and after retesting
Ito-Full, Ito-Core	Large-scale pooled Y2H data, with Core more reliable



- For DIP data, significant difference between CORE and Y2H
- Our own data is comparable to DIP-Y2H, even before the retesting
- Retest result is comparable to DIP-CORE

Thus, using three independent measures, we have confirmed that our high-confidence interaction data is indeed of very high quality, comparable to those produced by small-scale experiments, thus proving reproducibility is the key to filtering out false positives in Y2H screens.

To estimate the false positive rate of each dataset, we proceed as follows:

Let R be the reference correlation, which we take as the correlation for DIP-CORE, and let B be the background correlation between random pairs. Let x be the false positive rate, and hence $1 - x$ is the fraction of true positives in that dataset. Let G be the correlation of the given dataset, for which we would like to estimate its false positive rate.

Assuming that the false positives are random and that they are the only source of noise, we have

$$(1-x)R + xB = G, \text{ or } x = \frac{R-G}{R-B}$$

Using this formula we estimate that the (Y2H) false positive rate of our dataset is around 50% before retest and close to 0% afterwards. Ito-Core and Ito-Full are around 35% and 75%, respectively.

Bibliography

1. Watts, D.J. and S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, 1998. **393**(6684): p. 440-2.
2. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.
3. Schwikowski, B., P. Uetz, and S. Fields, *A network of protein-protein interactions in yeast.* Nat Biotechnol, 2000. **18**(12): p. 1257-61.
4. Li, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
5. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
6. Uetz, P., et al., *From ORFeomes to protein interaction maps in viruses.* Genome Res, 2004. **14**(10B): p. 2029-33.
7. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
8. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
9. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.
10. Bystroff, C. and D. Baker, *Prediction of local structure in proteins using a library of sequence-structure motifs.* J Mol Biol, 1998. **281**(3): p. 565-77.
11. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.
12. Teichmann, S.A., *The constraints protein-protein interactions place on sequence divergence.* J Mol Biol, 2002. **324**(3): p. 399-407.
13. Fraser, H.B., et al., *Evolutionary rate in the protein interaction network.* Science, 2002. **296**(5568): p. 750-2.

14. Mrowka, R., A. Patzak, and H. Herzel, *Is there a bias in proteome research?* Genome Res, 2001. **11**(12): p. 1971-3.

15. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.

16. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks.* Nature, 2000. **406**(6794): p. 378-82.

17. Maslov, S. and K. Sneppen, *Specificity and stability in topology of protein networks.* Science, 2002. **296**(5569): p. 910-3.

18. Berg, J. and M. Lassig, *Correlated random networks.* Phys Rev Lett, 2002. **89**(22): p. 228701.

19. Jenner, R.G., et al., *Kaposi's sarcoma-associated herpesvirus latent and lytic gene expression as revealed by DNA arrays.* J Virol, 2001. **75**(2): p. 891-902.

20. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. **9**(12): p. 3273-97.

21. Lehner, B. and A.G. Fraser, *A first-draft human protein-interaction map.* Genome Biol, 2004. **5**(9): p. R63.