Experimental and Computational Procedures for the Assessment of Protein Complexes on a Genome-wide Scale

Gabriel A. Musso,[†,‡] Zhaolei Zhang,[†,‡] and Andrew Emili*[,†,‡]

Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario, Canada M5S 3E1, and Department of Medical Genetics and Microbiology, University of Toronto, 1 Kings College Circle, Toronto, Ontario, Canada M5S 1A8

Received February 8, 2007

Contents

1. Introduction

Protein complexes, consisting of stable protein−protein interactions (PPIs), are ubiquitous and essential to the proper conduct of all eukaryotic functional pathways, serving to coordinate virtually every aspect of cellular biology.[1] Comprehensive determination of the entire network of protein interactions and highly associative protein units is useful in elucidating the mechanistic basis for complex biological processes and functionally characterizing interacting clusters of proteins. The term 'protein complex' has traditionally been used to describe heteromeric groups of tightly associated proteins that interact to form a unified cellular component such as the ribosome or proteasome. Yet, as large-scale interaction data has become increasingly available and global interaction networks discovered, the idea of the protein complex has evolved. That being to the notion of interconnected 'modules' consisting of groups of physically associated proteins functioning in a unified manner, although not necessarily with exclusive membership.[2] Consequently, researchers have begun to note heterogeneity in terms of both the apparent limited correlation of attributes such as gene coexpressions and the functional incongruence of putative members of certain protein complexes.[3−5] This has introduced a dichotomy in the interpretation of experimental datasets as some would define the protein complex as a stable macromolecule while others see it as a more dynamic, nonexclusive set of interacting proteins. Indeed, recent experimental evidence derived from genome-scale studies using yeast as a model system has begun to blur the heuristic boundaries that have historically been applied to define protein complexes as discrete biological articles.

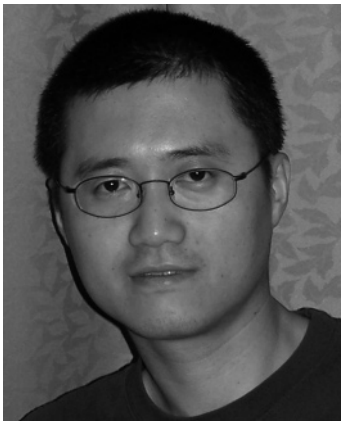
While the emergence of high-throughput interaction screening methods over the past 5 years now allows for more accurate and comprehensive elucidation of physical groupings of proteins in a systematic genome-wide manner, the rapidly increasing number of identified associations between gene products has ironically made it more difficult to clearly segregate proteins into discrete functional entities.[6]

As a result of this changing landscape, a most pressing challenge of the post-genomic era has become how to effectively integrate and accurately interpret the resulting global networks of physical associations so as to define functionally relevant modules that reflect ever-changing developmental cues, physiological signals, or disease-related maladaptive processes. Large-scale assessment of protein modules involves not only experimental analysis of protein associations but also computational evaluation of the reliability of the raw PPI data and thereafter algorithmic assignment of proteins into modules based on association data. As these modules are often of greatest interest to biologists, this review provides an overview of the integration of complementary experimental and computational methods that are currently available for the study of protein complexes on a genome scale. This review begins with an introduction and a description of the strengths and weaknesses of

* To whom correspondence should be addressed. E-mail: andrew.emili@utoronto.ca, Tel: 416-946-7281, Fax: 416-978-8287.
† Terrence Donnelly Centre for Cellular and Biomolecular Research.
‡ Department of Medical Genetics and Microbiology.

10.1021/cr0682857 CCC: $65.00
Published on Web 07/14/2007

Gabriel Musso is a Ph.D. student at the University of Toronto, Department of Medical Genetics and Microbiology, working out of the Donnelly Centre for Cellular and Biomolecular Research (CCBR) under the supervision of Zhaolei Zhang and Andrew Emili. He obtained his undergradaute degree in 2002 in the fields of Human Physiology and Physics and then pursued graduate studies at the Toronto Hospital for Sick Children, Department of Cardiovascular Surgery, in the discipline of Medical Science. After conferring his M.Sc. degree (2005), he obtained a postgraduate diploma in Bioinformatics from Seneca College (2005). His current thesis work focuses on analysis of both the physical and the genetic interactions of gene duplicates and subsequent determination of evolutionary patterns of duplicate gene retention.



Zhaolei Zhang is currently an Assistant Professor in the Banting & Best Department of Medical Research (BBDMR) and Department of Medical Genetics and Microbiology at the University of Toronto Faculty of Medicine. He received his B.S. degree from Nankai University in Tianjin, China, and Ph.D. degree in 2000 from the University of California at Berkeley, where he worked with Professor Sung-Hou Kim and Edward Berry on crystallographic studies of mitochondrial ubiquinone-cytochrome-*c* oxidoreductase (bc1 complex). From 2000 to 2004 he performed postdoctoral work with Professor Mark Gerstein at Yale University, Department of Molecular Biophysics and Biochemistry (MBB). There he investigated the properties and evolution of noncoding regions in the human genome, among other projects. In 2004 he started his own research group at the University of Toronto; his research laboratory is located in the Donnelly Centre for Cellular and Biomolecular Research (CCBR).

commonly used high-throughput interaction assays, both experimentally and computationally, highlighting notable papers. As protein interaction complexes are commonly algorithmically derived from lists of binary interaction data, methods excelling at resolving complexes and determining the presence or absence of single associations are discussed in the same context with the later sections of the review focusing on methods of computationally evaluating and deriving complexes from appropriate interaction data. Specifically, subsequent sections explain pertinent interaction

Andrew Emili is an associate professor in the Program in Proteomics and Bioinformatics at the Banting and Best Department of Medical Research and the Donnelly Centre for Cellular and Biomolecular Research at the University of Toronto, Canada. He is an expert in proteomics and molecular genetics. His research group performs genome-scale studies on the gene products of various model organisms using advanced proteomic technologies. Before joining the University of Toronto in 2000 he was a postdoctoral fellow at the Fred Hutchinson Research Center in Seattle, WA, with L. Hartwell, the noted geneticist and Nobel laureate (2001), and J. Yates III, a pioneer in protein mass spectrometry. A major research interest of his laboratory is the global investigation of protein expression, protein–protein interactions, and protein function.

validation methods as well as clustering algorithms and cluster-validation techniques used to evaluate derived protein complexes. Most of the methods mentioned here have been developed for data generation in budding yeast since it is currently the most extensively studied model organism for large-scale screens, although specific extensions to other organisms including human are mentioned.

2. Generation of Experimental Data

The goal of any proteome-scale association assay is high-quality interaction data. The *Saccharomyces cerevisiae* (budding yeast) proteome (which is relatively small in comparison to mammalian systems) has been investigated experimentally for over 30 years using low-throughput means.[7,8] Appropriately, the long-beheld gold standard in protein complexes in yeast was generally regarded to be the curated set stored in the Munich Information center for Protein Sequences (MIPS)[9] database, which contains experimentally well-characterized protein complexes generated through low-throughput assay. However, increasingly comprehensive lists of protein interactions were only recently generated with the application of high-throughput assays such as tandem affinity purification (TAP)[10] and yeast-2-hybrid (Y2H)[11] screening (see Figure 1b). Indeed, two recent global studies of yeast protein complexes published last year by Gavin et al.[12] and Krogan et al.[13] each predicted the existence of over 350 alternate groupings of proteins based on clustering (see section 3.2) of the physical interaction data. Yet while high-resolution interaction detection methods (Figure 1a) may avoid some of the problems, such as the often high false-positive and false-negative rates, associated with their high-throughput counterparts,[14,15] these low-throughput interaction methods are not practical for proteome-scale studies. For this reason, high-throughput interaction methods are emphasized in this section, while a brief overview of low-throughput assays is discussed in the context of data validation.

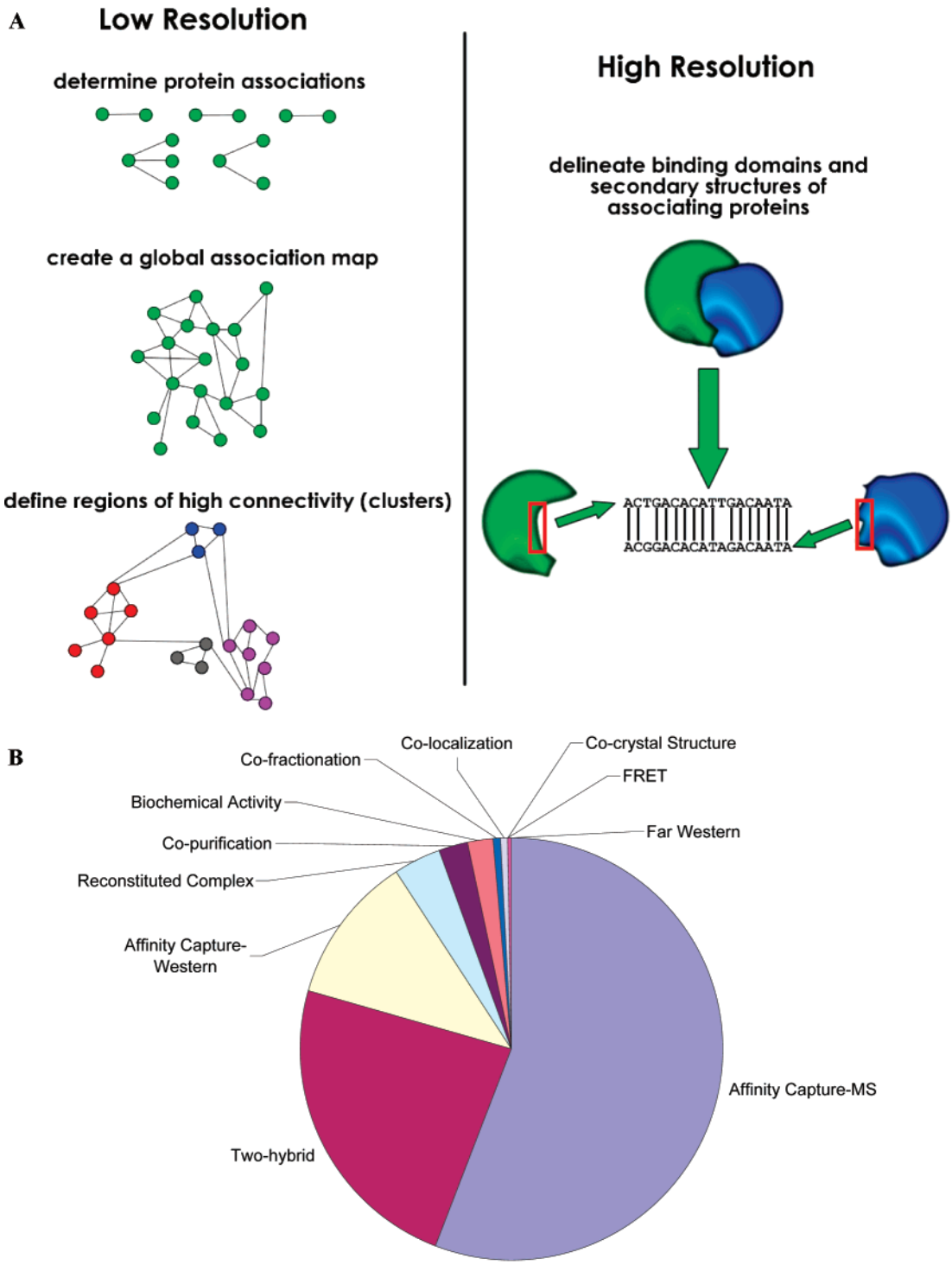


Figure 1. Basic overview of low- and high-resolution interaction surveying. (A) Low-resolution surveying (left) tends to begin with interaction data (typically in binary format), creates a cohesive interaction network map, and then assigns proteins to complexes through application of a clustering algorithm. This method, while often not as accurate as high-resolution mapping of macrocomplexes, can generally be applied to entire genomes. High-resolution surveying (right) is focused on determining conclusively the nature of protein associations, specifically through determination of associating secondary structure configurations. (B) Relative representation of interaction survey methods in the GRID database. The majority of representative interactions are the result of affinity capture-MS methods such as tandem affinity purification (TAP). The remaining majority are the result of yeast-2-hybrid studies.

2.1. Yeast-2-Hybrid

Y2H was first developed in the late 1980s[11] as a generalizable and highly sensitive method to screen for interactions among binary pairs of proteins and is still frequently used both as a first pass screening tool and for genome-scale exploratory studies today.[16−18] Despite several design variations since its inception which have resulted in improved assay efficiency,[19−22] the basic principle of the Y2H assay remains the same. That being that Y2H takes advantage of the fact that the process of transcriptional activation (and thus expression of a suitable reporter gene) depends on the tethering of two distinct protein domains to a target promoter: first, a DNA-binding domain (BD) that binds to the upstream DNA element and, second, an activation

domain (AD) that interacts with the general RNA polymerase machinery. In order to determine if an interaction occurs between two proteins, such as *x* and *y* (where *x* represents the bait protein and *y* the possible interactor or prey), protein *x* is expressed as a fusion to the DNA-binding domain, while the activation domain is likewise fused to protein *y*. If the re-engineered proteins are coexpressed and subsequently interact in the yeast nucleus, the jointly linked BD and AD will reconstitute an activator, leading to expression of a selectable reporter gene.[11] The presence or absence of a binary interaction can then be monitored on a large scale by screening thousands of strains for the activated expression of selectable markers and following the growth properties of viable yeast colonies. High-throughput adaptability can be further enhanced by mating ordered arrays of yeast strains encoding distinct bait and prey in a 96-well format.[23]

The first adaptation of the Y2H method to genome interaction mapping was reported for the T7 *E. coli* bacteriophage in which 25 interactions were identified among ~50 proteins.[24] The implications of this pioneering study were that the Y2H method could be applied to study interactions among the components encoded by a complex biological system or even an entire genome. In rapid succession the interaction networks of even more complex organisms were surveyed over the course of the next few years using the Y2H method. In 2001, two separate initiatives compiled Y2H-based global interaction maps in yeast,[25,26] noting a combined set of 4000 putative PPIs. In 2004, an initial first-pass proteome-wide map was reported for the nematode *Caenorhabditis elegans*,[27] the first study of its kind for a multicellular organism. Y2H screens have also been conducted in even more complex systems such as *Drosophila melanogaster* (fruit fly)[28,29] and most recently (although the data can be considered still preliminary) for human cells.[30,31]

A major advantage of Y2H is that reformation of the transcription factor complex used to detect interactions can occur when assayed proteins only transiently interact,[32] whereas comparable affinity-based purification methods (discussed below) have difficulty detecting transient interactions.[33] A major disadvantage of Y2H, however, is related to the often-elevated error rates. Analysis of large-scale datasets generated through Y2H tends to reveal low experimental overlap.[15,26,34] The most likely explanation for the lack of correlation between these two studies is a combination of both a high false-positive rate (estimated to be anywhere from 50%[15] to as to high as 90%[26,34]) and a false-negative rate, wherein most biologically relevant interactions are presumed to be missed. These artifacts stem in part from the overexpression and forced colocalization of the candidate proteins in the yeast nucleus, leading to nonphysiological context.[19] Consequently, while Y2H results are seen as a positive indication of a genuine protein interaction, the predictions benefit from additional supporting evidence.

Further limiting the applicability of Y2H in nonmodel organisms is its inability to survey interactions for gene products with incompletely defined coding sequences as an appropriate vector must be created for each query protein containing the associated gene and marker. This aspect limits implementation of comprehensive screens for mammalian systems where alternative splicing and incomplete knowledge of exons is common. Moreover, as Y2H is based on a binary interaction assay, it neglects interactions that involve three or more proteins,[34] which is precisely the hallmark of many, if not most, cellular protein complexes.

Due to the fact that interacting proteins must colocalize to the nucleus, Y2H has also traditionally not been useful for surveying interactions among integral membrane proteins. However, a specialized variant of Y2H, the so-called split-ubiquitin assay,[35] has been developed to tackle this missed opportunity and has shown increasing promise in recent years.[36] Briefly, one transmembrane (TM) domain containing protein is fused to the N-terminal half of ubiquitin, while the second TM protein is fused to the C-terminal half of ubiquitin and an adjoined transcription factor. Interaction of these proteins causes recognition of the complex by ubiquitin-recognizing proteases, thus releasing the transcription factor from its membrane anchor and thereby allowing subsequent activation of a reporter gene. Application of this method resulted in identification of nearly 2000 total interactions involving 536 TM-bearing proteins in yeast (131 of the 2000 interactions were deemed to be high quality based on a series of stringent criteria[36]).

Like most other Y2H-derived methods, the split-ubiquitin assay involves constitutively over-expressing the bait protein, often resulting in elevated (non-native) protein concentrations. Hence, the interactions captured by this approach may not occur at physiologic protein conditions, contributing to a high false-positive rate. Yet recent optimizations, such as integrating the tags into the target genome to achieve near-native expressions, may further the applicability of this assay for investigating the physical makeup of otherwise scantly characterized membrane-associated biochemical pathways.

Due to their ease of execution and scalability, Y2H-based binary assays remain prevalent in large-scale PPI surveys for many organisms despite their often high associated error rates. As computational methods for data evaluation improve (see section 3.1), biologists are becoming more adroit at reducing the number of false positives, thereby increasing the practical utility of such methods in establishing probable protein−protein interactions. However, the true pitfall when using Y2H data alone when trying to deduce the subunit composition of protein complexes is the generally high false-negative rate, which results in sparse representation of the overall biological networks of interactions and, consequently, poor assessment of discrete biological modules. Several of the large-scale Y2H screens have been shown to result in a network topology thought to be inconsistent with true biological systems[37] (for a primer on biological interaction networks and graph theory methods commonly applied to analyze them, see the Appendix). Therefore, for the purpose of deriving the molecular architecture of protein complexes, Y2H data alone is unfavorable.

2.2. Affinity Purification

In an effort to study protein complexes specifically and circumvent the inherent false-negative and false-positive rate of Y2H, affinity purification was developed for large-scale interaction surveys. The underlying concept behind affinity purification is a consequence of what had been observed in biochemical and coimmunoprecipitation studies for decades:[38] by selectively retrieving a protein of interest from a cell extract through use of a specific ligand or antibody, proteins stably bound to the query protein can usually be concomitantly retrieved. In affinity purification studies a universal epitope tag[39] is often systemically attached to the query proteins of interest which allows for routine bait capture along with any associated interactors via a single, well-defined, and often commercially available tag-specific

antibody. Proteins bound to the query protein are then usually identified through mass spectrometry using either traditional gel-based methods or gel-free tandem mass spectrometry procedures (the former offers qualitative information regarding subunit stoichiometry, while the latter provides superior sensitivity).

Affinity purification offers three distinct advantages over Y2H methods. First, only the query proteins require tagging, allowing novel interactions to be discovered between the baits and one or more poorly characterized proteins. Second, entire protein complexes can be captured during a single purification, as opposed to the binary interaction format employed by Y2H. Third, while the purification procedure can be tedious to scale up, the need to tag only one or two proteins in order to define a given complex reduces the number of experiments that need to be performed to achieve good proteomic coverage, compared to the multiple pairwise permutations ($n \times n$ experiments) required in a Y2H screen.

Using a systematic method of affinity purification coupled with mass spectrometry, the first interaction map for yeast was published in 2002.[40] Reflecting a substantive increase in proteome coverage, the group released an interaction map of higher density than previous comparable-scale studies consisting of 3617 putative protein−protein interactions for 493 tagged bait proteins. Importantly, the authors reported approximately 3-fold more interactions per protein which were curated in protein complex databases, suggesting a decreased false-negative rate.[40] However, while seemingly more accurate, the results of affinity purification studies have the unfortunate disadvantage of being biased against detection of low-abundance proteins with results dominated by higher-abundance bait proteins and often many spurious interactions resulting from common 'housekeeping' contaminants.

2.2.1. Tandem Affinity Purification

In an effort to increase the sensitivity of affinity purification to low-concentration proteins, the affinity purification process was further refined as tandem-affinity-purification.[10] The principle behind the TAP procedure is to retrieve proteins bound to epitope-tagged proteins of interest through two successive steps of affinity chromatography: first, generally via binding of the tagged protein to IgG beads and, second, via attachment to calmodulin (or an alternative affinity resin) beads.[10] Following the second elution the proteins (bait and interacting partners) are typically identified by mass spectrometry. The TAP interaction survey method is now recognized as having the best coverage and accuracy of experimental high-throughput interaction detection methods[15] and has the substantive advantage of detecting interactions among proteins assembled into protein complexes under near-native physiological conditions.

The first adaptation of the TAP method to large-scale protein complex characterization was performed in yeast and reported in 2001.[41] By tagging approximately 1700 proteins the authors were able to obtain data supporting 232 distinct functional interaction modules and provide hints as to the possible biological roles of 344 uncharacterized proteins based on physical association with proteins of known function.

In the following 4 years hundreds of studies were published identifying specific protein interactions and complexes not only in yeast but also in *E. coli*,[42] plant,[43] *drosophila*,[44] and human[45,46] (over 100 low-throughput studies as of 2004[47]). In 2006, two independent studies simultaneously published definitive, virtually comprehensive global surveys of stable soluble protein complexes for yeast.[12,13] Stringent data processing procedures were applied to the enormous raw datasets, seemingly eliminating false positives. Yet, despite the rigor and sheer scale of the two studies, with ~50% overlap in the total number of proteins detected (1304 in common out of the 1993 reported in Gavin et al.[12] study and the 2388 found in Krogan et al.[9]), initial cross-comparisons revealed a surprisingly modest overlap in the respective interactions (<25%; see Figure 2). Aside from minor differences in the respective screening procedures, the most likely cause for this seeming shortfall stems from differences in the computational algorithms used to ascertain the most likely protein interactions and interaction clusters,[48] thus illustrating the impact of the increasingly sophisticated analytical methods used to interpret genome-scale interaction data (discussed in more detail in section 3). It should however be noted that although TAP is specifically designed for complex resolution, final determination of complexes has depended on algorithmic interpretation of determined confidence scores between pairs of interactors; thus, the potential exists to misrepresent experimentally determined complexes. These issues are discussed in more detail in subsequent computational sections. One other caveat to TAP screening is the functional interference potentially caused by introduction of the epitope tag or selection marker, which may perturb protein folding/function and mRNA stability/regulation. Despite being relatively innocuous, initial reports suggested that a tag may impair function inasmuch as 18% of all targeted proteins.[41] It is worth noting, however, that one-third[12] to nearly one-half[13] of interacting proteins were identified as untagged preys for other tagged proteins and thus could still be surveyed.

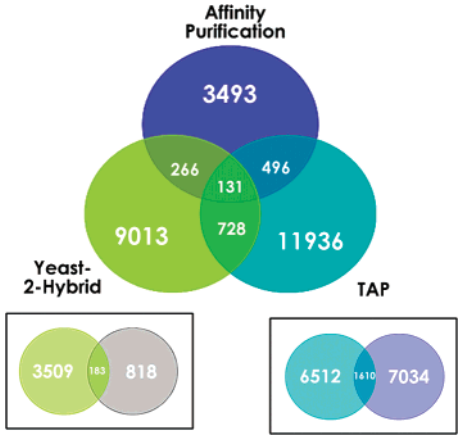


Figure 2. Overlap of high-throughput interaction study results. The larger Venn diagram shows the overlap in protein interaction data resulting from three independent methods: tandem affinity purification (TAP), yeast-2-hybrid, and affinity purification combined with mass spectrometry. All interaction data was obtained from the GRID database. The inset Venn diagrams show the overlap in interactions of the two largest single studies for each respective method (TAP and yeast-2-hybrid). In the case of yeast-2-hybrid inset studies are those published by Ito et al. and Uetz et al. For TAP, the two inset studies are by Gavin et al. and Krogan et al.

2.3. Genome-Scale Fluorescence Tagging and Microscopy

Much as the name implies, fluorescence tagging involves addition of a detectible tag to a protein of interest through

Table 1. Interactions per Publication for Commonly Applied Interaction Detection Methods[a]

method	ints/pub
affinity capture-MS	128.64
two hybrid	11.48
FRET	6.14
copurification	4.78
affinity capture-western	3.39
biochemical activity	2.92
affinity capture-RNA	2.62
cofractionation	2.59
colocalization	2.08
cocrystal structure	2.06
reconstituted complex	2.06
far western	1.64

[a] Affinity capture-MS averages more interactions per publication than other high-throughput methods such as two hybrid. Two-hybrid assay is still commonly applied in low-throughput format to test interaction of specific proteins, resulting in its low interaction per publication average. Methods appearing in the table below two-hybrid typically require far too much experimental time to be adapted to global experimental assay.

either antibody binding or epitope tagging. Traditionally, antibodies selected against a target protein (primary antibodies) are reacted with dye-coupled secondary antibodies (recognizing the primary reagents), allowing cellular localization in fixed cells to subsequently be monitored through microscopy.[49] In 2002, the first large-scale localization survey[50] was performed in yeast using indirect immunofluorescence of 2744 proteins tagged with a universal epitope marker, illustrating the primary subcellular localization of about one-half of the protein complement of yeast. By recombining query proteins with a recognizable tag, the tedious task of having to create antibodies specific for each protein of interest is avoided.

Originally obtained from jellyfish, green fluorescent protein (GFP) is likely the most commonly used fluorescent marker for expression and localization assays. Through ligation and transformation, genomic sequence encoding GFP can be attached to genes of interest which then in turn become detectible through microscopy upon expression. As recombination technology has progressed and entire genome sequences become known, proteome-scale GFP tagging has become feasible. In 2003, a virtually complete GFP-tagging survey of the yeast proteome (4156 proteins in total) was reported,[51] again avoiding use of specific antibodies altogether. A unified pattern emerged—proteins existing within the same transcriptional module (groups of transcriptionally coregulated proteins reasoned to be functionally related[52,53]) exhibited similar localization patterns. Further, it was noted that proteins known to interact physically or functionally were far more likely to exist in the same cellular compartment,[51] although colocalization alone (even as demonstrated through high-resolution localization assays) does not provide evidence of a direct physical interaction. However, while the colocalization of two proteins may not strictly indicate that they interact or coexist within a protein complex,[54] it is often implicative of a functional relationship at some level, while conversely, the lack of colocalization is used as evidence of noninteractors.[55]

2.4. New and Arising Methods of Interaction Detection: FRET and LUMIER

Förster Resonance energy transfer (FRET)[56,57] (Table 1) represents a highly specialized means of interaction detection, similar in principle to fluorescence tagging and recently adapted to the study of protein complexes.[58] During FRET two proteins are tagged and coexpressed with variants of GFP. When in immediate proximity (<80 Å) energy transfer between the juxtaposed fluorescent molecules occurs and the resulting emission spectra can be monitored.[59] Due to the exceptional resolution and sensitivity of FRET, the method can be used to examine spatial relationships of interacting pairs of proteins within a cell.[60] A major advantage of the FRET approach is in the ability to study transient PPIs within living cells[60] as even very brief interactions result in quantifiable energy transfers. The technique has also been adapted to examining the activation states of tyrosine kinases[61] in living mammalian cells and formation of pheromone-responsive protein complexes[58] in yeast.

A characteristic pitfall of FRET, however, is the substantial amount of optimization that must be dedicated to reducing photobleaching in the experimental output (for a more in-depth expository, see Sekar and Periasamy[62]). Hence, for the time being FRET remains a specialized low-throughput assay, although it is conceivable that FRET could be adapted to high-throughput format as both the core technology and associated optics continue to improve.

Another new method for examining broader sets of inducible protein−protein interactions is a highly sensitive luminescence-based mammalian interactome mapping (LUMIER)[63] system, which was recently introduced for a mammalian cell setting. In the LUMIER approach a protein of interest is fused with a Renilla luciferase reporter enzyme, while potential binding partners are coexpressed as fusions to a FLAG epitope-tag.[64] Immunoprecipitation is performed in a parallel, multiwell manner using anti-FLAG antibody specific for the preys and the luminescence associated with each pull down then quantified via spectrophotometry. One advantage of the LUMIER method is that it is well suited to examination of temporal kinetics for dynamic protein interaction networks. Accordingly, the procedure was used to explore interactions induced among components of a transforming growth factor beta (TGF-β) pathway following activation of signaling in the context of an immortalized human cell line.[63] In the process of performing a total of 12 000 binary experiments the authors examined the molecular dynamics of pathway activation and thereby managed to link TGF-β to a novel kinase cascade. However, one notable drawback of the LUMIER method is that, like Y2H methods, non-native overexpression of the test proteins is required. Moreover, the complexity and amount of experimental data generated for even a modest-scale pilot study examining a single pathway suggests that scaling up to genome-scale studies will be challenging. Thus, while this method has shown considerable promise for targeted PPI surveys, further refinements are necessary before it is generally applied in a more comprehensive manner.

2.5. Computational Methods for Detecting Protein–Protein Interactions

The rapidly mounting reams of experimentally generated interaction data now available for diverse model organisms such as yeast, *E. coil*, fly, worm, and also increasingly for human poses an ever growing challenge to bioinformaticians both for quality control and biological inference. Appropriate bioinformatic analyses are essential for properly interpreting high-throughput experimental results so as to amass, process, and distill relevant interactions. Fortunately, the substantive

breadth of innovative computational methods and tools introduced over the past few years greatly facilitate the efficient gathering and accurate assessment of large-scale interaction datasets, allowing researchers to define functionally relevant protein clusters and extended interaction modules. Computational biologists are likewise becoming more adept at combining and extrapolating interaction data to provide more profound biological insights. This section outlines the various computational approaches that have been applied toward generation of binary interaction data, ultimately leading to the algorithmic delineation of the architecture of protein complexes.

2.5.1. Orthology Mapping

While there exists copious amounts of interaction data for single-cell organisms (i.e., yeast), there remains relatively sparse experimental data for mammalian species such as mouse or human. Although the physiology is vastly different, many aspects of cellular function (and by extension protein complexes) are well conserved across divergent species.[65] For this reason, the concept of cross-species interaction mapping has been investigated extensively.[66−69] That is, the findings derived for simpler models, such as single cell yeast, can be extrapolated to higher multicellular eukaryotes based on the concept of orthology.

Conceptually, conserved binary interactions can be mapped by determining orthologous gene pairs across species and then transferring analogous interactions reported in one species to another (these conserved protein interactions are known as interologs).[70] For example, if proteins *a* and *b* have been reported to interact in yeast while human cells express the putative orthologs, *c* and *d*, an interaction between *c* and *d* can be predicted and examined.

The traditional method of determining orthology is through high sequence similarity (i.e., alignment of two genes in separate genomes indicates that they are more like each other than any member of the opposing genomes). In fact, results indicate that interactions can be reliably transferred using sequence identities above 80%.[71] This logic is employed by the InParanoid[72] database, which has been used to infer protein interactions between orthologous gene products for several dozen species.[73,74]

One can achieve increased accuracy in the assignment of orthology and predicted protein interactions by comparing sequences and combining pools of interaction data obtained from multiple species at once as theoretically the integration process reduces noise from errors for a particular species.[75] Such is the logic underlying the Clusters of Orthologous Groups (COG) database[76] that uses multiple-species phylogenetic mapping to determine groups of orthologous proteins. Further applying the notion that increasing the number of species being compared bolsters confidence in ortholog assignment, the STRING database[77] expands on COG data by examining additional species and then further extrapolates physical interaction data toward even newly sequenced species.

Through orthology mapping interactions derived for organisms with less complex and experimentally more tractable proteomes can be transferred to a mammalian setting. Indeed, large-scale PPI networks based on orthology inferences have since been drawn for human cells[78,79] and even been applied to the study of human disease states.[80,81]

While orthologous gene products are often similar in sequence, there is no guarantee that the respective proteins indeed perform identical cellular functions or in fact keep the same protein interactions. In addition, it may be difficult or impossible to ascertain accurate orthology relationships after gene family expansions, where one-to-many genetic associations are more prevalent than a simplified direct one-to-one relationship. Furthermore, other factors such as expression and subcellular localization patterns could easily change across species, making the physical interaction of certain combinations of proteins impossible regardless of the strength of a perceived association. It is perhaps for these reasons that many such inferred interactions have been shown to have low overall predictive accuracy.[82]

Recent advances in orthology mapping have begun incorporating available functional information from multiple experimental sources[83] in response to the criticism that even valid orthologs need not be functionally identical across species; however, extrapolated interactions can still be considered speculatory. Orthology mapping, specifically to detect the conservation of entire protein complexes, requires more sophisticated algorithms capable of incorporating expression, localization, and other functional data in order to more exactly determine conserved protein interactions. Such algorithms do not exist yet to our knowledge but likely will be developed within the next few years as such supporting genomic data becomes available for an increasing number of species. Until then, predictions generated through orthology mapping are not preferred over validated experimental data.

2.5.2. Sequence- and Structure-Based Associations

Interpretation of coding sequence to infer to domain architecture is commonly used to determine the potential for protein interactions.[84] In the late 1990s a multiple-species genomic sequence comparison method was developed to predict physical interactions and other functional associations between proteins.[85] This method is based on the frequent observation that two peptide chains existing in one species are often encoded as disjoint domains in a single gene in another species. For example, the *E. coli* GyrA and GyrB gene products were initially thought to form a complex (a prediction confirmed by experimentation in 2005[42]) since each of these factors has high similarity with a domain encoded by a singly larger protein (topoisomerase II) in yeast.[85] Since the information gained from one organism helps to predict an interaction between two proteins in another species, this method is referred to as the *Rosetta Stone* approach, in analogy to the method used in deciphering ancient languages. However, the limited generality of cross-species relationships limits this approach's utility for comprehensive projections.

On a more promising note, more precise elucidation of protein interaction domains may increase the power of such procedures. One recent study[86] proposed that there are a finite number of conceivable structural configurations of physical interactions that may occur between two proteins of any species. This value was estimated to be approximately 10 000 with the interaction domains of roughly 2000 already defined.[86] With the full exposition of the exact structures and sequences of interaction domains it might be possible to infer an interaction between any pair of proteins based on amino-acid sequence identity alone. Accurate predictions of the potential of proteins to interact can even be made based on as little as 60% sequence similarity to known PPI binding interfaces.[87] In 2003, all possible pairwise permutations of

yeast proteins were evaluated for plausible combinations of well-established interdomain associations present in a dimer database,[88] resulting in over 7000 predicted PPIs, one-half of which were supported by alternate evidence derived from other methods. One advantage of this computational approach is that it is intrinsically not biased against characterization of low-abundance proteins,[88] unlike affinity-based experimental procedures.

Forecasting the composition of entire complexes represents an increase in convolution over binary interaction prediction and accordingly an increase in computation time, decreasing its applicability to large-scale complex prediction. Also, as was the shortcoming of orthology mapping, this method gives no indication about whether a given interaction truly occurs in vivo as potential interactors may be temporally or spatially uncoupled. However, the recent overlay of structural protein characteristics with known protein-interaction networks has highlighted new evolutionary properties of highly associative proteins and shown great potential in future interaction network analysis.[89] This area can be expected to expand greatly in the coming years.

2.5.3. Text Mining

As individual researchers typically gather information for proteins of interest through examining publications, text mining remains among the most established methods of gathering PPI data. However, since the number of publications available today is expanding explosively for virtually every subspecialty of the biological sciences, comprehensive text interpretation has become too demanding. To accommodate this data overload, computer programs have been developed to systematically process and parse out interaction data from large bodies of published literature in an automated manner.

Algorithms that scan the literature for PPI have two tasks: first, to recognize conclusively instances of mentioned proteins (which is challenging because proteins often have multiple names and abbreviations) and, second, to define the biophysical context in which these proteins are being discussed.[90] Due to the complexity of the English language and to the varied nature of protein interactions themselves, defining context is no trivial task. More importantly, the algorithm must be able to accurately distinguish a genuine interaction from a coincidental occurrence, which may present with nearly the same lexical syntax. Effective algorithms must be trained to recognize appropriate English sentence features (including verbal form, presence/absence of a noun) and scored against manual curation to evaluate performance before finally being used to parse large bodies of literature.[91] Recent computational approaches have integrated information from sentences both preceding and following the mention of protein/gene names to improve the accuracy of the general approach[92,93] (for a more in-depth review, see Jensen et al.[90] and Hirschman et al.[94]). Today there are several publicly available software packages for performing automated literature mining of protein interactions (MEDSYNDIKATE[95] and CONAN[96]).

Text mining can be an effective data collection method for several reasons. First, the data collected is often from multiple experimental sources, resulting in a collective set of interactions less subject to any specific experimental bias. Second, as the data retrieved by text mining algorithms is vastly beyond what is published in any single (even high-throughput) experiment, there is increased potential for cross-validation. For example, there are over 80 000 yeast-specific protein interactions in the last release of the GRID[97] database. If one filters and reduces this set to only the most accurate 1% of interactions (suitable evaluation methods are discussed in more detail below), an even larger subset of putative interactions is generated than obtained for either the Krogan et al.[13] or Gavin et al.[12] published datasets. For these reasons, text mining in combination with manual curation has been used to populate public databases such as preBIND[98] and GRID,[97] which are invaluable for computational biologists as they tend to represent the largest sources of PPI data for every species. The accuracy of interactions housed within these databases significantly increases, however, when literature retrieval algorithms are coupled with manual curation,[98] as experts can comb through and remove as many spurious interactions as possible. Interaction networks created from literature-mined protein interactions also exhibit topology similar to networks generated by high-throughput screening alone, although with better coverage[99] (see the Appendix for a brief discussion of graph theory and network topology).

As for the study of mammalian protein interactions, a potential disadvantage of text mining is the prevalence of literature for commonly studied proteins, such as those associated with cancer or other widely studied processes or diseases. The increased representation skews the resulting interaction map. Consequently, the accuracy of text mining algorithms also improves as more experimental data (and hence publications) is generated for a given organism.

3. Computational Approaches To Assessing Putative PPI Networks

Yeast has developed into the gold-standard testing ground for evaluating large-scale computational and experimental approaches to interaction surveys. Benchmarking the methods described above has indicated often-impressive efficacy within a unicellular model setting but more importantly exhibited a potential to be adapted to the more challenging frontier of mammalian interaction assay. One of the most striking lessons to be learned for large-scale interaction assays is the importance of vigilant data management. While experimental and computational high-throughput methods are constantly increasing in accuracy and scope, the importance of stringent data validation, especially when interpreting protein interaction networks, remains paramount. Generating experimental data is therefore only the first of several steps in understanding a biological interaction network. Accordingly, while the preceding sections have focused on creation of the data pieces, the remainder of this review will be focused on properly resolving these into an informative puzzle.

One intriguing aspect of resolved interaction networks has been the prevalence of cross-connections between various protein modules as projected by certain high-throughput screens in yeast,[6,100] which suggests a preponderance of cross-talk among biological systems. Increasingly sophisticated computational procedures are now adept at determining both the core components and more transient members of the functional modules that underlie protein interactions,[12,14,100] ultimately aiming to comprehend the mesh-like structure of the interactome. Yet while high-throughput experimental and computational methods are poised to provide a major advance for systems biologists seeking to understand how integration of PPI networks occurs on a global scale, use of computa-

tional algorithms to classify proteins into one or more finite complexes based on sparse interaction data can pose a lack of precision. Many researchers are interested in appraising the properties of a relatively small number of constitutively bound, functionally unified stable protein complexes. For this reason, the decision of which computational methods are subsequently used to evaluate experimental data depends as much on an investigator's own definition of the protein complex as it does on the nature of the interactome itself. Regardless of one's stance, however, computational delineation of protein complexes tends to begin first with evaluation of raw the input PPI, followed by derivation of interconnected protein clusters, and last some form of evaluation of clustering accuracy. Each of these steps is discussed in detail below.

3.1. Evaluating Raw Protein–Protein Interaction Data

Every method of interaction discovery or retrieval has an associated false-positive and false-negative rate. Appropriately, one of the most important steps in characterizing PPI networks is determination of how extensively the network may be over- or under-representative of a biological system of particular interest. Removal of potentially spurious data is a priority for large-scale assay, evidenced by the fact that typically only the most confident 10−20% is published.[12,13] Unfortunately, this can introduce a paradox: how does one determine if putative new interactions are a true reflection of reality and not simply the result of poor coverage when there is scant validated information?

There are two general approaches to interaction data validation. One can either rely on properties of the interaction data alone or evaluate interactions based on secondary, and potentially circumstantial, supporting evidence. If evaluating networks based on data obtained from one sole experimental source (i.e., large-scale Y2H or a TAP assay), the simplest method of assigning confidence scores is often based on the reported experimental properties. For example, interactions generated through identifications of the subunits of purified protein complexes using mass spectrometry are generally given confidence scores based on certainty and reproducibility of the initial protein identifications. However, evaluating interactions based on broader network properties may be more appropriate for studies that ultimately aim to understand the network as a whole. In this way, interactions can be scored or assigned a probability based on how well they suit other functional attributes or the properties of known interaction graphs. In 2006, Gavin et al.[12] released a large-scale list of protein complexes predictions for yeast based on clustering of their experimentally recorded networks of PPI data. Their scoring system, known as the socio-affinity-index (SAI), was based on a "spoke" pairing model originally proposed by Bader and Hogue.[14] This model states that a protein will bind to its fellow complex members like the spokes of a wheel with a bait protein having direct binary interactions with all prey proteins. In terms of application, SAI assigns a score for pairs of putatively interacting proteins based on the fraction of times they copurify. This method controlled the total number of proteins retrieved for each bait protein, penalizing promiscuous (i.e., most likely false-positive) binding partners. SAI values were then translated to binary interaction confidence scores based on comparison with a set of gold-standard reference PPI obtained from the MIPS database. Appropriate confidence score cut offs are usually determined through receiver-operating-characteristic (ROC) curve analysis (for a description of the ROC process, see Hanley and McNeil[101]).

Evaluation based solely on experimentally generated interactions (internal evaluation) is advantageous in that it avoids biases introduced by potentially nonrepresentative external data. However, any systematic bias introduced as a result of the experimental technique will, in turn, be reflected in the generated interaction network. In contrast, data integration during interaction evaluation (herein referred to as external evaluation) relies on publicly accepted knowledge to evaluate experimental data. In this context, application of machine learning can be useful. While a review of machine learning is well beyond the scope of this text, the basic concept is that an algorithm can be trained to assign probabilities to new events based on its previous exposure to well-accepted, similar events. For protein interaction data, a machine-learning algorithm can be trained to recognize patterns based on properties of well-known interactors (i.e., gold standards found in public curated databases as well as other high confidence experimental results). The algorithm is then used to examine the properties of novel interactors, implicated by either experimental or computational means, and in doing so determine the likelihood of that a given interaction is spurious. For example, in a recent publication Krogan et al.[13] used a machine learning algorithm trained on the experimental properties (reproducibility of raw PPI scores, etc.) of well-accepted protein interactions detected among a small number of experimentally validated gold-standard protein complexes stored in the MIPS[9] database and then applied the same algorithm to assign confidence scores for every novel interaction in the experimental dataset. After determining an appropriate confidence score cutoff through ROC-curve analysis, the authors obtained a high-confidence dataset containing 7123 interactions among 2708 proteins from their original set of over 70 000 interactions.

Although Krogan et al.[13] used MIPS complexes to influence confidence score assignment, studies have demonstrated that evaluation based on more speculatory evidence can be effective as well. As mentioned briefly above (section 2.3), interacting proteins are known to be enriched for certain functional attributes as compared to noninteractors. Meta-analyses of data generated for yeast indicates that interacting proteins are far more likely to coexist in specific cellular compartments,[51] share patterns of transcriptional activation,[3,4] and share other correlated properties.[102] For this reason, similar annotation in the Gene Ontology (GO) database,[103] correlated expression patterns,[3,4] or colocalization[51] can be used to influence evaluation of experimental data, and combination of supplementary data of several types can effectively be tied to machine learning to integrate multiple interaction datasets.[104−106]

3.1.1. Low-Throughput Methods Commonly Applied in Data Validation

Although characteristically only existing for small populations of proteins, interaction data generated through higher resolution assays are typically very accurate and therefore can be used as an independent benchmark for high-throughput experimental datasets. For example, the MIPS complexes that are commonly used in both interaction and cluster benchmarking are often characterized using methods discussed herein.

While generally applied to solving the structures of single proteins, an increasing number of complexes including the

very large proteasome[107] and even the complete ribosome[108] have been deduced using X-ray crystallography. Large-scale application of X-ray crystallography to protein complex analysis, however, is still limited due to the difficulty associated with producing substantive amounts of highly purified heteromeric proteins as well as complications in crystallization.[109]

For those complexes that are not easily determined through X-ray crystallography (such as membrane proteins which are notoriously difficult to crystallize), single-particle cryo-electron microscopy (cryo-EM) represents an attractive alternative technique. Cryo-EM generates 3-dimensional projections of complexes of interest by combining multiple 2-dimensional 'slice' images of the dried sample. This represents a decrease in expense over X-ray crystallography but also an increase in overall analysis time. Due to the nature of the images obtained, any PPI predicted by cryo-EM must be combined with either another experimental approach[110] or rigorous low-throughput validation methods.[111]

Data from nuclear magnetic resonance (NMR) has also been used to examine the structure and functional relationships of proteins for nearly three decades[112,113] and importantly facilitate characterization of at least 14 protein complexes in the last 3 years[113] (for a review, see Bonvin et al.[113] and Mittermaier and Kay[114]). NMR methodology detects spectral changes associated with conformational rearrangement of backbone residues as a result of multiprotein binding.[115] However, relatively high concentrations of sample are needed, and data collection can take days to weeks.

3.2. Clustering of Interaction Data

After obtaining interaction data (either experimentally or computationally), the next challenge then becomes that of assigning proteins into individual complexes. Computational partitioning of interaction networks into highly connected clusters has been used to impressive effect in large-scale studies of yeast[12,13,116] and human.[81,117]

Although varied, clustering algorithms usually define subgroups of proteins that exhibit higher similarity among themselves than with other subgroups.[118] In defining interaction clusters, which are posited to represent protein complexes, there are several algorithms that can be used, the selection of which will depend on the nature of the desired outcome. Some algorithms produce individual (exclusive) clusters with nonshared members; others will allow shared members between clusters. Nonexclusivity (i.e., clusters can have shared members) can be viewed as being more biologically accurate as many proteins show promiscuity in terms of complex membership. However complexes with nonshared members ease postanalysis of results and facilitate functional categorizations. Additionally, some algorithms are capable of incorporating biological or functional data.

The lack of overlap[48] in corresponding yeast protein interactions reported by Krogan et al.[13] and Gavin et al.[12] points to the importance of selecting a standardized computational assessment procedure. Recent follow-up studies[119] (also Pu, S.; Wodak, S. Personal communication) demonstrate that application of a unified clustering method results in similar clusters for both datasets. Moreover, an additional caveat is that while disparate algorithms can decipher alternate interconnected groups of proteins, the results serve only as an approximation of the actual physical complexes present within the cell. Many of the algorithms described operate based on properties of graph theory; for a brief description of graph theory in biology, see the Appendix.

3.2.1. Clustering Algorithms

k-means is recognized as being one of the simplest clustering algorithms in application today. For a set of X clusters, X centroid values will be determined equally covering the range of the inputted set. Data points are then individually assigned to the centroid that they are closest in value to. Unfortunately, in order to use *k*-means clustering, the number of clusters must be anticipated in advance. This represents a major disadvantage when studying novel interactomes as the number of complexes present is impossible to predetermine.

Commonly depicted using a dendrogram, hierarchical clustering is famously used in biology to classify species based on phenotypic or phylogenetic properties. More recently, hierarchical clustering has been applied to PPI data, finding proteins with highly similar expression patterns.[120] There are several variations of hierarchical clustering; however, all those commonly applied in interaction cluster analysis consist of the same steps. Each node in the set being analyzed begins the process as its own cluster. From there, similarity between any two nodes in terms of properties such as minimal path length[121] is computed using one of many different measures (i.e., Spearman's rank). The two nodes that are the most similar are moved into the same cluster, and distances to all other nodes are recomputed. This is continued until the entire tree structure has been established. Examining proteins grouped together at one level of the hierarchy allows one to draw finite protein clusters.

In a purely computational study Krause et al.[122] applied three variations of hierarchical clustering to a yeast affinity-purification dataset, ultimately concluding that more interaction data is required for an accurate complexosome description. Several years later Gavin et al.[12] built on Krause's results and used a similar approach to draw their PPI clusters based on experimental data in the yeast proteome. This nonexclusive method also integrated functional data when deriving clusters and was able to group proteins into either stable protein 'modules' comprising the functional core of unified protein complexes or extended promiscuous associators, designations the established authors felt to be truly indicative of the state of the complexosome.

One method that is gaining increased attention for PPI clustering is the Markov clustering algorithm (MCL).[123] Once a network graph of proteins has been generated (Figure 3) random 'walks' are created in silico (wherein nodes are picked at random and a predetermined number of PPI "edges" is traversed). Through an iterative process of many such walks the algorithm splits the proteins into exclusive groups based on the relative flow across highly traversed regions (high connectivity indicates clusters). In a recent comparison of biologically applied clustering algorithms[124] MCL was shown to be remarkably resilient to spurious graph perturbations. Appropriately, MCL was used to describe many novel protein complexes within the yeast proteome based on large-scale TAP experimental data.[13]

Another algorithm, known as MCODE,[125] used for detecting protein complexes among PPI networks similarly divides interaction data into clusters based on regions of high connectivity. This algorithm is freely available as a plug-in for the Cytoscape[126] software package, which allows for ready viewing of the agglomerative results (Figure 3).

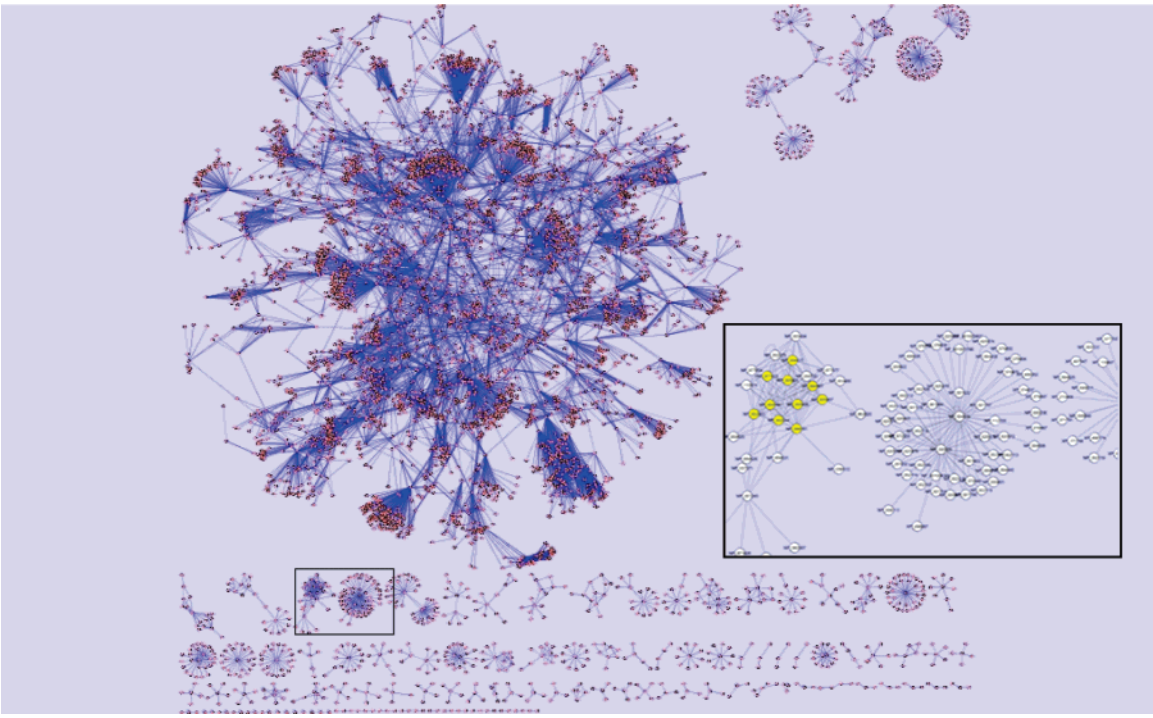


Figure 3. Visualization of a protein interaction network. The Cytoscape software package was used to display the first 10 000 interactions in the human cancer protein interaction map. Display was created using Cytoscape's organic layout option. The large cluster in the middle of the figure represents the largest connected subgroup of the interaction network, while connections below and to the above right of the graph represent other nonconnected units. (Inset) Representation of one cluster obtained using the MCODE algorithm. Cluster members are highlighted.

3.2.2. Evaluation of Predicted Protein Clusters

Determining the accuracy of computationally derived protein complexes is more challenging than establishing the accuracy of the corresponding binary interactions. This is partly due to the fact that there are varying definitions of the protein complex itself, while protein interactions are more clearly defined. Also, it is feasible to establish negative gold standards for protein interactions, suggesting which proteins might not interact,[55] while it is generally far more difficult to say with certainty which proteins do not exist in the same complex. If there were a well-established set of reference protein complexes, assessing the accuracy of members would be a more realistic issue. Yet such gold-standard complexes are infrequent, especially for mammalian systems. Last, whereas binary protein interactions have been studied on a large scale for almost a decade, high-throughput protein complex determination studies only began a few years ago. Therefore, the resulting state of knowledge of protein complexes is less than comprehensive.

As mentioned briefly above, the MIPS[9] database currently contains 68 stable complexes (217 subcategorized complexes) characterized in yeast consisting of 1297 nonexclusive subunits and is widely regarded as being the most comprehensive list of protein complexes available. This reference set was used in the machine-learning algorithm of Krogan et al.[13] to assign confidence scores to novel protein interactions and complexes. Gavin et al.[12] likewise compared their computationally derived complexes with reference to the MIPS dataset to determine clustering accuracy and found their experimental dataset to have 73% coverage. However, even the highly curated MIPS database is suspected to be biased and contain errors,[55] and it is debatable if experimental divergence from MIPS data is a sign of decreased accuracy or gain of knowledge.

Aside from comparisons to known complexes, another approach in protein complex validation is to examine overlapping functional characterization using protein/gene annotations in databases such as GO,[103] OMIM,[127] or KEGG[128] or examine similarity of the patterns of gene expression. This stems from the logic that interacting proteins are by definition functionally linked and the experimental observation that proteins of similar function show similar expression patterns.[129−131] However, recent evidence has shown that the pathways and functional modules of cells may be subject to much more cross-talk and intermember promiscuity than previously thought,[6,132] and therefore, functional- and expression-based interpretations can lead to oversimplification.

Ultimately, clusters are perhaps best modeled after a biologist's own sense of what constitutes a functional module. Some authors prefer to use gold standards as an objective measure, which they believe to represent protein complexes and hence score PPI data accordingly. As previously mentioned, a common framework for firmly establishing the nature of protein modules would serve to unify the community and potentially increase the body of known complexes; however, as of yet no such initiatives exist.

4. Human Complexosome: A Case Study

With the broad array of both experimental and computational methods as well as the lack of consensus in defining the nature of the protein complex, definitive protocols for large-scale complex detection can be difficult to establish. For this reason a working example of how protein complexes might be elucidated in a human setting can help to illustrate the issues and caveats at hand.

Earlier this year the task was laid upon our lab to draft a preliminary map of protein complexes in human. As experimental methods for mammalian complexosome study are still being evaluated and computational studies require far less of an investment both of time and money, the logical approach is to attempt to generate this list computationally. The only de novo computational method for detecting interactions is through interolog mapping, which, for reasons mentioned above, is not preferable to actual experimental data. As low-throughput studies have been undertaken in human cells for decades, literature mining of interactions should generate both a larger and a more accurate list of protein interactions in human than orthology mapping. While several literature-mining software packages may be appropriate to apply here, public databases have already begun manually curating human interaction data (more below). The resulting interactions have higher fidelity than whatever a lone research group may retrieve. If the focus of study had been an organism of less interest, predefined algorithms would have been applied.

The Human Protein Reference Database (HPRD)[133] contains over 36 000 freely available, literature-mined, curated protein interactions in human, making it the largest public repository of human interaction data.[134] Importantly though, each interaction is listed along with a publication number referencing both the paper and method used to detect the interaction. This information can be used to re-examine the original published experimental results and assign confidence scores based on experimental techniques. A small collection of interactions chosen randomly (say 5% of the data) can be manually checked against the listed publication to determine if the literature-mining algorithm obtained a true interaction. By evaluating how often the algorithm reported the interaction properly, a rough indication of the accuracy of the literature-mining process is obtained.

If the dataset is found to be of sufficient quality, the next step in obtaining human protein clusters is to assign confidence scores to these 36 000 interactions. Since both internal and external evaluation methods have their merit, a good approach would be to apply both to interaction scoring. By internal evaluation methods we refer to those computer algorithms that can assign confidence scores by examining the topology and structure properties of the interaction networks. Examples of these methods are the social association index (SAI) as used by Gavin et al.[12] (mentioned in section 3.1). External methods include those algorithms that utilize properties of individual proteins or genes to infer or validate pairwise interactions. For example, the presence of protein domains and phenotypic characterization of a small gold-standard subset of protein interactions (for example, those experimentally determined through NMR or cocrystallization) could be used to train a machine-learning algorithm to deduce the confidence scores in the remaining interactions. Final confidence scores can then be an amalgam of those obtained from internal and external evaluation methods and an appropriate score cutoff determined through ROC analysis.

Once the interactions are scored, the next step is algorithmic clustering followed by independent evaluation of the putative protein clusters. MCL is an attractive algorithm in this context since it is capable of automatically incorporating the previously determined confidence scores as edge weights during the process of assigning cluster membership. As there is currently no widely accepted gold-standard set of human protein complexes, evaluation of the accuracy of the clustering exercise would depend on the properties of protein clusters determined in other organisms. For example, clusters can be examined for coexpression and colocalization among the predicted subunits to obtain some estimate of the expected biological enrichment. The most preferable means of validation, however, would be some form of in vivo experimental testing of a small subset of randomly selected protein clusters, preferably using a reliable biophysical method such as TAP or Y2H.

The process described above is obviously an oversimplification of a definitive pipeline of defining the multitude of human complexes but illustrates a successful integration of key methods discussed in this review.

5. Conclusions and Perspectives

While there exists no single method best suited to characterize protein complexes in vivo, modern proteomic procedures and modeling algorithms are steadily increasing in accuracy as is the amount of publicly available interaction data collected for important organisms. Over the next 5 years the landscape of high-throughput interaction surveying can be expected to change dramatically (just as it has quickly expanded from 5 years ago) due to the substantive improvement in both existing and innovative new experimental and computational techniques. However, as high-throughput interaction screening inevitably moves away from simpler model systems, such as yeast, and inexorably toward mammalian organisms, a new battery of challenges will be encountered. Among these are the increased magnitude and dynamic nature of protein interactions and complexes in the multicellular milieu, more complex patterns of regulation, more extensive post-translation modifications of complex subunits in various tissues and cell types, and more complicated stress and environmental control mechanisms. A resulting computational difficulty to overcome will be delineating complex membership among the far more numerous set of protein interactions.

Systematic functional assessment of human gene products steadily increases[135,136] since completion of sequencing of the genome.[137,138] Over the past several years the total number of estimated functional genes in the genome has decreased to the currently accepted number of 21−23,000, which are seemingly within reach for a battery of powerful new survey technologies. Yet at the same time the magnitude of interactions per protein has progressively increased, suggesting that the increased complexity of higher organisms does not lie in the genetic code per se but rather in the physical and functional associations of proteins. Additional confounders, such as alternative splicing, will likewise cause a consequent increased reliance on computational methods for benchmarking until newly specialized experimental techniques appear for mammalian systems. Accordingly, breakthroughs in techniques capable of describing the labyrinth-like landscape of the human interactome are eagerly anticipated.

6. Glossary

centroid	geometrically, the centroid is the point of intersection of hypergeometric planes of an object (i.e., the exact center of a triangle)
complexosome	coined from similarity to proteome and interactome, the complexosome describes all protein complexes existing within a cell

cytoscape	software package specialized for interaction network analysis maintained by a public effort (http://www.cytoscape.org)
dendrogram	tree-style diagram used to depict elements grouped together by a clustering algorithm
edge	literally, the lines connecting members of a network graph. In the context of protein association diagrams, edges correspond to physical interactions between proteins
gold standard	in the context of protein complexes, the most trustworthy set of complexes from which to benchmark experimental and computational approaches
gene ontology (GO) database	hierarchically structured repository of gene characterizations for multiple organisms. The three main categories of characterization are localization, biological process, and molecular function (http://www.geneontology.org)
interaction domain	the term 'interaction domain' has several interpretations within computational biology. As it was applied in this manuscript, it corresponds to a given section of the genomic or amino-acid sequence encoding a secondary structure known to interact with other similarly described secondary structures
interactome	all protein interactions within a cell
interolog	protein interaction conserved across species
minimal path length	shortest number of edges required to connect two nodes on a network graph
node	members of network graphs; in this case, proteins
ortholog	genes in two species are considered orthologous if they were both inherited from a common precursor gene in an ancestral species
proteome	entire protein complement of a cell
topology	while topology can be used to describe a specific branch of geometry, in this context it refers to the structure of connections in a network graph, be they scale free or random in nature
transforming growth factor beta (TGF-β)	the TGF-β superfamiliy of proteins has been implicated in such processes as cell growth, division, and differentiation. Since the TGF-β pathway has a distinct effect on programmed cell death (apoptosis), it has been studied extensively in the development of cancer.

7. Acknowledgments

We thank Ruth Isserlin for graciously reviewing this manuscript. This study was supported in part by funds from Genome Canada through the Ontario Genomics Institute.

8. Appendix

8.1. Introduction to Graph Theory

Graphs with proteins or genes depicted as nodes and interactions as the edges between them are frequently used to represent both protein and genetic interaction networks. Using graphs of this type to depict interaction networks allows analysis by a certain set of mathematical formulas, those pertaining to graph theory.

Many naturally occurring graphs exhibit a topology in which the majority of constituent nodes have only a few associations and just a few nodes have many associations (i.e., the number of associations per protein follows a power-law distribution, with corresponding graphs referred to as being scale-free). Scale-free graphs exhibit 'small-world' characteristics as most nodes can be linked to one another by following a short path. Models of this type are generally noted in real-world networks; in fact, the so-called 'small-world' problem[139] was originally described when Stanley Milgram noticed shorter than expected path lengths among social networks (later the basis for the famous 'six degrees of separation' hypothesis). It is tempting to believe that scale-free topology exists in biological networks as it would offer greater protection against random deletions than other types of associative graph structures.

Whether the PPI network truly follows a scale-free pattern remains controversial with compelling evidence presented both supporting[140,141] and negating[142,143] the claim. Regardless, proteins with higher connectivity in the network (often called 'hubs') tend to be more essential to cell function.[140,144] Similarly, graph properties such as betweenness and closeness (literally measuring the relative proximities of two nodes on a network graph) are often used to describe how functionally related two proteins or genes are.[145,146] There are several freely available software packages used in the analysis of PPI networks, the most common of which are Pajek[123] and Cytoscape[126] with Cytoscape having many useful plug-ins for biological analysis.

9. References

(1) Alberts, B. *Cell* **1998**, *92*, 291−4.
(2) Hartwell, L. H.; Hopfield, J. J.; Leibler, S.; Murray, A. W. *Nature* **1999**, *402*, C47−52.
(3) Ge, H.; Liu, Z.; Church, G. M.; Vidal, M. *Nat. Genet.* **2001**, *29*, 482−6.
(4) Jansen, R.; Greenbaum, D.; Gerstein, M. *Genome Res.* **2002**, *12*, 37−46.
(5) Carmi, S.; Levanon, E. Y.; Havlin, S.; Eisenberg, E. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **2006**, *73*, 031909.
(6) Batada, N. N.; Reguly, T.; Breitkreutz, A.; Boucher, L.; Breitkreutz, B. J.; Hurst, L. D.; Tyers, M. *PLoS Biol.* **2006**, *4*, 1−12.
(7) Trewyn, R. W.; Nakamura, K. D.; O'Connor, M. L.; Parks, L. W. *Biochim. Biophys. Acta* **1973**, *327*, 336−44.
(8) Penninckx, M. *Eur. J. Biochem.* **1975**, *58*, 533−8.
(9) Mewes, H. W.; Amid, C.; Arnold, R.; Frishman, D.; Guldener, U.; Mannhaupt, G.; Munsterkotter, M.; Pagel, P.; Strack, N.; Stumpflen, V.; Warfsmann, J.; Ruepp, A. *Nucleic Acids Res.* **2004**, *32*, D41−4.
(10) Puig, O.; Caspary, F.; Rigaut, G.; Rutz, B.; Bouveret, E.; Bragado-Nilsson, E.; Wilm, M.; Seraphin, B. *Methods* **2001**, *24*, 218−29.
(11) Fields, S.; Song, O. *Nature* **1989**, *340*, 245−6.
(12) Gavin, A. C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dumpelfeld, B.; Edelmann, A.; Heurtier, M. A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A. M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; Superti-Furga, G. *Nature* **2006**, *440*, 631−6.
(13) Krogan, N. J.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Ignatchenko, A.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A. P.; Punna, T.; Peregrin-Alvarez, J. M.; Shales, M.; Zhang, X.; Davey, M.; Robinson, M. D.; Paccanaro, A.; Bray, J. E.; Sheung, A.; Beattie, B.; Richards, D. P.; Canadien, V.; Lalev, A.; Mena, F.; Wong, P.; Starostine, A.; Canete, M. M.; Vlasblom, J.; Wu, S.; Orsi, C.; Collins, S. R.; Chandran, S.; Haw, R.; Rilstone, J. J.; Gandi, K.; Thompson, N. J.; Musso, G.; St Onge, P.; Ghanny, S.; Lam, M. H.; Butland, G.; Altaf-Ul, A. M.; Kanaya, S.; Shilatifard, A.; O'Shea, E.; Weissman, J. S.; Ingles, C. J.; Hughes, T. R.; Parkinson, J.; Gerstein, M.; Wodak, S. J.; Emili, A.; Greenblatt, J. F. *Nature* **2006**, *440*, 637−43.
(14) Bader, G. D.; Hogue, C. W. *Nat. Biotechnol.* **2002**, *20*, 991−7.
(15) von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P. *Nature* **2002**, *417*, 399−403.
(16) Landazuri, M. O.; Vara-Vega, A.; Viton, M.; Cuevas, Y.; del Peso, L. *Biochem Biophys. Res. Commun.* **2006**, *351*, 313−20.
(17) Kim, J. S.; Rho, B.; Lee, T. H.; Lee, J. M.; Kim, S. J.; Park, J. H. *Biochem Biophys. Res. Commun.* **2006**, *351*, 253−8.
(18) Solaz-Fuster, M. C.; Gimeno-Alcaniz, J. V.; Casado, M.; Sanz, P. *Cell Signal.* **2006**, *18*, 1702−12.
(19) Vidal, M.; Legrain, P. *Nucleic Acids Res.* **1999**, *27*, 919−29.
(20) Bauer, A.; Kuster, B. *Eur. J. Biochem.* **2003**, *270*, 570−8.
(21) Colland, F.; Jacq, X.; Trouplin, V.; Mougin, C.; Groizeleau, C.; Hamburger, A.; Meil, A.; Wojcik, J.; Legrain, P.; Gauthier, J. M. *Genome Res.* **2004**, *14*, 1324−32.

(22) Obrdlik, P.; El-Bakkoury, M.; Hamacher, T.; Cappellaro, C.; Vilarino, C.; Fleischer, C.; Ellerbrok, H.; Kamuzinzi, R.; Ledent, V.; Blaudez, D.; Sanders, D.; Revuelta, J. L.; Boles, E.; Andre, B.; Frommer, W. B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12242−7.

(23) Walhout, A. J.; Vidal, M. *Methods* **2001**, *24*, 297−306.

(24) Bartel, P. L.; Roecklein, J. A.; SenGupta, D.; Fields, S. *Nat. Genet.* **1996**, *12*, 72−7.

(25) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; QuReshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. *Nature* **2000**, *403*, 623−7.

(26) Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4569−74.

(27) Li, S.; Armstrong, C. M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P. O.; Han, J. D.; Chesneau, A.; Hao, T.; Goldberg, D. S.; Li, N.; Martinez, M.; Rual, J. F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S. L.; Zhang, L. V.; Berriz, G. F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H. W.; Elewa, A.; Baumgartner, B.; Rose, D. J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S. E.; Saxton, W. M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K. C.; Harper, J. W.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Vidal, M. *Science* **2004**, *303*, 540−3.

(28) Stanyon, C. A.; Liu, G.; Mangiola, B. A.; Patel, N.; Giot, L.; Kuang, B.; Zhang, H.; Zhong, J.; Finley, R. L., Jr. *Genome Biol.* **2004**, *5*, R96.

(29) Formstecher, E.; Aresta, S.; Collura, V.; Hamburger, A.; Meil, A.; Trehin, A.; Reverdy, C.; Betin, V.; Maire, S.; Brun, C.; Jacq, B.; Arpin, M.; Bellaiche, Y.; Bellusci, S.; Benaroch, P.; Bornens, M.; Chanet, R.; Chavrier, P.; Delattre, O.; Doye, V.; Fehon, R.; Faye, G.; Galli, T.; Girault, J. A.; Goud, B.; de Gunzburg, J.; Johannes, L.; Junier, M. P.; Mirouse, V.; Mukherjee, A.; Papadopoulo, D.; Perez, F.; Plessis, A.; Rosse, C.; Saule, S.; Stoppa-Lyonnet, D.; Vincent, A.; White, M.; Legrain, P.; Wojcik, J.; Camonis, J.; Daviet, L. *Genome Res.* **2005**, *15*, 376−84.

(30) Rual, J. F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D. S.; Zhang, L. V.; Wong, S. L.; Franklin, G.; Li, S.; Albala, J. S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R. S.; Vandenhaute, J.; Zoghbi, H. Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M. E.; Hill, D. E.; Roth, F. P.; Vidal, M. *Nature* **2005**, *437*, 1173−8.

(31) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksoz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H.; Wanker, E. E. *Cell* **2005**, *122*, 957−68.

(32) Causier, B.; Davies, B. *Plant Mol. Biol.* **2002**, *50*, 855−70.

(33) Bauch, A.; Superti-Furga, G. *Immunol. Rev.* **2006**, *210*, 187−207.

(34) Ito, T.; Ota, K.; Kubota, H.; Yamaguchi, Y.; Chiba, T.; Sakuraba, K.; Yoshida, M. *Mol. Cell. Proteomics* **2002**, *1*, 561−6.

(35) Johnsson, N.; Varshavsky, A. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 10340−4.

(36) Miller, J. P.; Lo, R. S.; Ben-Hur, A.; Desmarais, C.; Stagljar, I.; Noble, W. S.; Fields, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12123−8.

(37) Wagner, A. *Mol. Biol. Evol.* **2001**, *18*, 1283−92.

(38) Phizicky, E. M.; Fields, S. *Microbiol. Rev.* **1995**, *59*, 94−123.

(39) Fritze, C. E.; Anderson, T. R. *Methods Enzymol.* **2000**, *327*, 3−16.

(40) Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Hogue, C. W.; Figeys, D.; Tyers, M. *Nature* **2002**, *415*, 180−3.

(41) Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* **2002**, *415*, 141−7.

(42) Butland, G.; Peregrin-Alvarez, J. M.; Li, J.; Yang, W.; Yang, X.; Canadien, V.; Starostine, A.; Richards, D.; Beattie, B.; Krogan, N.; Davey, M.; Parkinson, J.; Greenblatt, J.; Emili, A. *Nature* **2005**, *433*, 531−7.

(43) Rubio, V.; Shen, Y.; Saijo, Y.; Liu, Y.; Gusmaroli, G.; Dinesh-Kumar, S. P.; Deng, X. W. *Plant J.* **2005**, *41*, 767−78.

(44) Veraksa, A.; Bauer, A.; Artavanis-Tsakonas, S. *Dev. Dyn.* **2005**, *232*, 827−34.

(45) Bouwmeester, T.; Bauch, A.; Ruffner, H.; Angrand, P. O.; Bergamini, G.; Croughton, K.; Cruciat, C.; Eberhard, D.; Gagneur, J.; Ghidelli, S.; Hopf, C.; Huhse, B.; Mangano, R.; Michon, A. M.; Schirle, M.; Schlegl, J.; Schwab, M.; Stein, M. A.; Bauer, A.; Casari, G.; Drewes, G.; Gavin, A. C.; Jackson, D. B.; Joberty, G.; Neubauer, G.; Rick, J.; Kuster, B.; Superti-Furga, G. *Nat. Cell Biol.* **2004**, *6*, 97−105.

(46) Brajenovic, M.; Joberty, G.; Kuster, B.; Bouwmeester, T.; Drewes, G. *J. Biol. Chem.* **2004**, *279*, 12804−11.

(47) Dziembowski, A.; Seraphin, B. *FEBS Lett.* **2004**, *556*, 1−6.

(48) Goll, J.; Uetz, P. *Genome Biol.* **2006**, *7*, 223.

(49) Dagger, F.; Dunia, I.; Hernandez, A. G.; Pradel, L. A.; Benedetti, E. L. *Mol. Biol. Rep.* **1988**, *13*, 197−206.

(50) Kumar, A.; Agarwal, S.; Heyman, J. A.; Matson, S.; Heidtman, M.; Piccirillo, S.; Umansky, L.; Drawid, A.; Jansen, R.; Liu, Y.; Cheung, K. H.; Miller, P.; Gerstein, M.; Roeder, G. S.; Snyder, M. *Genes Dev.* **2002**, *16*, 707−19.

(51) Huh, W. K.; Falvo, J. V.; Gerke, L. C.; Carroll, A. S.; Howson, R. W.; Weissman, J. S.; O'Shea, E. K. *Nature* **2003**, *425*, 686−91.

(52) Holstege, F. C.; Jennings, E. G.; Wyrick, J. J.; Lee, T. I.; Hengartner, C. J.; Green, M. R.; Golub, T. R.; Lander, E. S.; Young, R. A. *Cell* **1998**, *95*, 717−28.

(53) Ihmels, J.; Friedlander, G.; Bergmann, S.; Sarig, O.; Ziv, Y.; Barkai, N. *Nat. Genet.* **2002**, *31*, 370−7.

(54) Wallrabe, H.; Periasamy, A. *Curr. Opin. Biotechnol.* **2005**, *16*, 19−27.

(55) Jansen, R.; Gerstein, M. *Curr. Opin. Microbiol.* **2004**, *7*, 535−45.

(56) Miyawaki, A.; Llopis, J.; Heim, R.; McCaffery, J. M.; Adams, J. A.; Ikura, M.; Tsien, R. Y. *Nature* **1997**, *388*, 882−7.

(57) Siegel, M. M.; Tabei, K.; Kagan, M. Z.; Vlahov, I. R.; Hileman, R. E.; Linhardt, R. J. *J. Mass Spectrom.* **1997**, *32*, 760−72.

(58) Raicu, V.; Jansma, D. B.; Miller, R. J.; Friesen, J. D. *Biochem. J.* **2005**, *385*, 265−77.

(59) Majoul, I.; Straub, M.; Duden, R.; Hell, S. W.; Soling, H. D. *J. Biotechnol.* **2002**, *82*, 267−77.

(60) Day, R. N.; Schaufele, F. *Mol. Endocrinol.* **2005**, *19*, 1675−86.

(61) Ting, A. Y.; Kain, K. H.; Klemke, R. L.; Tsien, R. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 15003−8.

(62) Sekar, R. B.; Periasamy, A. *J. Cell. Biol.* **2003**, *160*, 629−33.

(63) Barrios-Rodiles, M.; Brown, K. R.; Ozdamar, B.; Bose, R.; Liu, Z.; Donovan, R. S.; Shinjo, F.; Liu, Y.; Dembowy, J.; Taylor, I. W.; Luga, V.; Przulj, N.; Robinson, M.; Suzuki, H.; Hayashizaki, Y.; Jurisica, I.; Wrana, J. L. *Science* **2005**, *307*, 1621−5.

(64) Witzgall, R.; O'Leary, E.; Bonventre, J. V. *Anal. Biochem.* **1994**, *223*, 291−8.

(65) Wells, L.; Fridovich-Keil, J. L. *SAAS Bull. Biochem. Biotechnol.* **1996**, *9*, 83−8.

(66) Persico, M.; Ceol, A.; Gavrila, C.; Hoffmann, R.; Florio, A.; Cesareni, G. *BMC Bioinformatics* **2005**, *6* (Suppl 4), S21.

(67) Koyuturk, M.; Kim, Y.; Subramaniam, S.; Szpankowski, W.; Grama, A. *J. Comput. Biol.* **2006**, *13*, 1299−322.

(68) Boulton, S. J.; Gartner, A.; Reboul, J.; Vaglio, P.; Dyson, N.; Hill, D. E.; Vidal, M. *Science* **2002**, *295*, 127−31.

(69) Kelley, B. P.; Yuan, B.; Lewitter, F.; Sharan, R.; Stockwell, B. R.; Ideker, T. *Nucleic Acids Res.* **2004**, *32*, W83−8.

(70) Matthews, L. R.; Vaglio, P.; Reboul, J.; Ge, H.; Davis, B. P.; Garrels, J.; Vincent, S.; Vidal, M. *Genome Res.* **2001**, *11*, 2120−6.

(71) Yu, H.; Luscombe, N. M.; Lu, H. X.; Zhu, X.; Xia, Y.; Han, J. D.; Bertin, N.; Chung, S.; Vidal, M.; Gerstein, M. *Genome Res.* **2004**, *14*, 1107−18.

(72) O'Brien, K. P.; Remm, M.; Sonnhammer, E. L. *Nucleic Acids Res.* **2005**, *33*, D476−80.

(73) Liang, Z.; Xu, M.; Teng, M.; Niu, L. *BMC Bioinformatics* **2006**, *7*, 457.

(74) Zhong, W.; Sternberg, P. W. *Science* **2006**, *311*, 1481−4.

(75) Sharan, R.; Suthram, S.; Kelley, R. M.; Kuhn, T.; McCuine, S.; Uetz, P.; Sittler, T.; Karp, R. M.; Ideker, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1974−9.

(76) Tatusov, R. L.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Kiryutin, B.; Koonin, E. V.; Krylov, D. M.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; Smirnov, S.; Sverdlov, A. V.; Vasudevan, S.; Wolf, Y. I.; Yin, J. J.; Natale, D. A. *BMC Bioinformatics* **2003**, *4*, 41.

(77) von Mering, C.; Jensen, L. J.; Kuhn, M.; Chaffron, S.; Doerks, T.; Kruger, B.; Snel, B.; Bork, P. *Nucleic Acids Res.* **2007**, *35*, D358−62.

(78) Lehner, B.; Fraser, A. G. *Genome Biol.* **2004**, *5*, R63.
(79) Rhodes, D. R.; Tomlins, S. A.; Varambally, S.; Mahavisno, V.; Barrette, T.; Kalyana-Sundaram, S.; Ghosh, D.; Pandey, A.; Chinnaiyan, A. M. *Nat. Biotechnol.* **2005**, *23*, 951−9.
(80) Oti, M.; Snel, B.; Huynen, M. A.; Brunner, H. G. *J. Med. Genet.* **2006**, *43*, 691−8.
(81) Jonsson, P. F.; Bates, P. A. *Bioinformatics* **2006**.
(82) Mika, S.; Rost, B. *PLoS Comput. Biol.* **2006**, *2*, e79.
(83) Jonsson, P. F.; Cavanna, T.; Zicha, D.; Bates, P. A. *BMC Bioinformatics* **2006**, *7*, 2.
(84) Lin, K.; Zhu, L.; Zhang, D. Y. *Bioinformatics* **2006**, *22*, 2081−6.
(85) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. *Science* **1999**, *285*, 751−3.
(86) Aloy, P.; Russell, R. B. *Nat. Biotechnol.* **2004**, *22*, 1317−21.
(87) Aloy, P.; Pichaud, M.; Russell, R. B. *Curr. Opin. Struct. Biol.* **2005**, *15*, 15−22.
(88) Lu, L.; Arakaki, A. K.; Lu, H.; Skolnick, J. *Genome Res.* **2003**, *13*, 1146−54.
(89) Kim, P. M.; Lu, L. J.; Xia, Y.; Gerstein, M. B. *Science* **2006**, *314*, 1938−41.
(90) Jensen, L. J.; Saric, J.; Bork, P. *Nat. Rev. Genet.* **2006**, *7*, 119−29.
(91) Sugiyama, K.; Hatano, K.; Yoshikawa, M.; Uemura, S. *Genome Informatics* **2003**, *14*, 699−700.
(92) Ding, J.; Berleant, D.; Nettleton, D.; Wurtele, E. *Pac. Symp. Biocomput.* **2002**, 326−37.
(93) Hahn, U.; Romacker, M.; Schulz, S. *Pac. Symp. Biocomput.* **2002**, 338−49.
(94) Hirschman, L.; Park, J. C.; Tsujii, J.; Wong, L.; Wu, C. H. *Bioinformatics* **2002**, *18*, 1553−61.
(95) Hahn, U.; Romacker, M.; Schulz, S. *Int. J. Med. Inform.* **2002**, *67*, 63−74.
(96) Malik, R.; Franke, L.; Siebes, A. *Bioinformatics* **2006**, *22*, 2151−7.
(97) Breitkreutz, B. J.; Stark, C.; Tyers, M. *Genome Biol.* **2003**, *4*, R23.
(98) Donaldson, I.; Martin, J.; de Bruijn, B.; Wolting, C.; Lay, V.; Tuekam, B.; Zhang, S.; Baskin, B.; Bader, G. D.; Michalickova, K.; Pawson, T.; Hogue, C. W. *BMC Bioinformatics* **2003**, *4*, 11.
(99) Reguly, T.; Breitkreutz, A.; Boucher, L.; Breitkreutz, B. J.; Hon, G. C.; Myers, C. L.; Parsons, A.; Friesen, H.; Oughtred, R.; Tong, A.; Stark, C.; Ho, Y.; Botstein, D.; Andrews, B.; Boone, C.; Troyanskya, O. G.; Ideker, T.; Dolinski, K.; Batada, N. N.; Tyers, M. *J. Biol.* **2006**, *5*, 11.
(100) Sprinzak, E.; Altuvia, Y.; Margalit, H. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 14718−23.
(101) Hanley, J. A.; McNeil, B. J. *Radiology* **1982**, *143*, 29−36.
(102) Letovsky, S.; Kasif, S. *Bioinformatics* **2003**, *19* (Suppl 1), i197−204.
(103) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* **2000**, *25*, 25−9.
(104) Lee, I.; Date, S. V.; Adai, A. T.; Marcotte, E. M. *Science* **2004**, *306*, 1555−8.
(105) Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N. J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J. F.; Gerstein, M. *Science* **2003**, *302*, 449−53.
(106) Lu, L. J.; Xia, Y.; Paccanaro, A.; Yu, H.; Gerstein, M. *Genome Res.* **2005**, *15*, 945−53.
(107) Lowe, J.; Stock, D.; Jap, B.; Zwickl, P.; Baumeister, W.; Huber, R. *Science* **1995**, *268*, 533−9.
(108) Yusupov, M. M.; Yusupova, G. Z.; Baucom, A.; Lieberman, K.; Earnest, T. N.; Cate, J. H.; Noller, H. F. *Science* **2001**, *292*, 883−96.
(109) Russell, R. B.; Alber, F.; Aloy, P.; Davis, F. P.; Korkin, D.; Pichaud, M.; Topf, M.; Sali, A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 313−24.
(110) Gao, H.; Sengupta, J.; Valle, M.; Korostelev, A.; Eswar, N.; Stagg, S. M.; Van Roey, P.; Agrawal, R. K.; Harvey, S. C.; Sali, A.; Chapman, M. S.; Frank, J. *Cell* **2003**, *113*, 789−801.
(111) Stark, H.; Luhrmann, R. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 435−57.
(112) Wuthrich, K.; Wagner, G.; Richarz, R.; Perkins, S. J. *Biochemistry* **1978**, *17*, 2253−63.
(113) Bonvin, A. M.; Boelens, R.; Kaptein, R. *Curr. Opin. Chem. Biol.* **2005**, *9*, 501−8.
(114) Mittermaier, A.; Kay, L. E. *Science* **2006**, *312*, 224−8.
(115) Palmer, A. G., III; Kroenke, C. D.; Loria, J. P. *Methods Enzymol.* **2001**, *339*, 204−38.
(116) Schwikowski, B.; Uetz, P.; Fields, S. *Nat. Biotechnol.* **2000**, *18*, 1257−61.
(117) Ramani, A. K.; Bunescu, R. C.; Mooney, R. J.; Marcotte, E. M. *Genome Biol.* **2005**, *6*, R40.
(118) Backer, E.; Jain, A. *IEEE Trans. Pattern Anal. Mach. Intell.* **1981**, *PAMI-3*, 66−75.
(119) Collins, S. R.; Kemmeren, P.; Zhao, X. C.; Greenblatt, J. F.; Spencer, F.; Holstege, F. C.; Weissman, J. S.; Krogan, N. J. *Mol. Cell. Proteomics* **2007**, *6*, 439−50.
(120) Krogan, N. J.; Peng, W. T.; Cagney, G.; Robinson, M. D.; Haw, R.; Zhong, G.; Guo, X.; Zhang, X.; Canadien, V.; Richards, D. P.; Beattie, B. K.; Lalev, A.; Zhang, W.; Davierwala, A. P.; Mnaimneh, S.; Starostine, A.; Tikuisis, A. P.; Grigull, J.; Datta, N.; Bray, J. E.; Hughes, T. R.; Emili, A.; Greenblatt, J. F. *Mol. Cell* **2004**, *13*, 225−39.
(121) Arnau, V.; Mars, S.; Marin, I. *Bioinformatics* **2005**, *21*, 364−78.
(122) Krause, R.; von Mering, C.; Bork, P. *Bioinformatics* **2003**, *19*, 1901−8.
(123) Batagelj, V.; Mrvar, A *Connections* **1998**, *21*, 47−57.
(124) Brohee, S.; van Helden, J. *BMC Bioinformatics* **2006**, *7*, 488.
(125) Bader, G. D.; Hogue, C. W. *BMC Bioinformatics* **2003**, *4*, 2.
(126) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Res.* **2003**, *13*, 2498−504.
(127) Hamosh, A.; Scott, A. F.; Amberger, J.; Bocchini, C.; Valle, D.; McKusick, V. A. *Nucleic Acids Res.* **2002**, *30*, 52−5.
(128) Kanehisa, M. *Novartis Found. Symp.* **2002**, *247*, 91−101; discussion 101−3, 119−28, 244−52.
(129) Niehrs, C.; Pollet, N. *Nature* **1999**, *402*, 483−7.
(130) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863−8.
(131) Grigoriev, A. *Nucleic Acids Res.* **2001**, *29*, 3513−9.
(132) Krause, R.; von Mering, C.; Bork, P.; Dandekar, T. *Bioessays* **2004**, *26*, 1333−43.
(133) Mishra, G. R.; Suresh, M.; Kumaran, K.; Kannabiran, N.; Suresh, S.; Bala, P.; Shivakumar, K.; Anuradha, N.; Reddy, R.; Raghavan, T. M.; Menon, S.; Hanumanthu, G.; Gupta, M.; Upendran, S.; Gupta, S.; Mahesh, M.; Jacob, B.; Mathew, P.; Chatterjee, P.; Arun, K. S.; Sharma, S.; Chandrika, K. N.; Deshpande, N.; Palvankar, K.; Raghavnath, R.; Krishnakanth, R.; Karathia, H.; Rekha, B.; Nayak, R.; Vishnupriya, G.; Kumar, H. G.; Nagini, M.; Kumar, G. S.; Jose, R.; Deepthi, P.; Mohan, S. S.; Gandhi, T. K.; Harsha, H. C.; Deshpande, K. S.; Sarker, M.; Prasad, T. S.; Pandey, A. *Nucleic Acids Res.* **2006**, *34*, D411−4.
(134) Mathivanan, S.; Periaswamy, B.; Gandhi, T.; Kandasamy, K.; Suresh, S.; Mohmood, R.; Ramachandra, Y.; Pandey, A. *BMC Bioinformatics* **2006**, *7* (Suppl 5), S19.
(135) Collins, F. S.; Lander, E. S.; Rogers, J.; Waterston, R. H. *Nature* **2004**, *431*, 931−45.
(136) She, X.; Jiang, Z.; Clark, R. A.; Liu, G.; Cheng, Z.; Tuzun, E.; Church, D. M.; Sutton, G.; Halpern, A. L.; Eichler, E. E. *Nature* **2004**, *431*, 927−30.
(137) McPherson, J. D.; Marra, M.; Hillier, L.; Waterston, R. H.; Chinwalla, A.; Wallis, J.; Sekhon, M.; Wylie, K.; Mardis, E. R.; Wilson, R. K.; Fulton, R.; Kucaba, T. A.; Wagner-McPherson, C.; Barbazuk, W. B.; Gregory, S. G.; Humphray, S. J.; French, L.; Evans, R. S.; Bethel, G.; Whittaker, A.; Holden, J. L.; McCann, O. T.; Dunham, A.; Soderlund, C.; Scott, C. E.; Bentley, D. R.; Schuler, G.; Chen, H. C.; Jang, W.; Green, E. D.; Idol, J. R.; Maduro, V. V.; Montgomery, K. T.; Lee, E.; Miller, A.; Emerling, S.; Kucherlapati; Gibbs, R.; Scherer, S.; Gorrell, J. H.; Sodergren, E.; Clerc-Blankenburg, K.; Tabor, P.; Naylor, S.; Garcia, D.; de Jong, P. J.; Catanese, J. J.; Nowak, N.; Osoegawa, K.; Qin, S.; Rowen, L.; Madan, A.; Dors, M.; Hood, L.; Trask, B.; Friedman, C.; Massa, H.; Cheung, V. G.; Kirsch, I. R.; Reid, T.; Yonescu, R.; Weissenbach, J.; Bruls, T.; Heilig, R.; Branscomb, E.; Olsen, A.; Doggett, N.; Cheng, J. F.; Hawkins, T.; Myers, R. M.; Shang, J.; Ramirez, L.; Schmutz, J.; Velasquez, O.; Dixon, K.; Stone, N. E.; Cox, D. R.; Haussler, D.; Kent, W. J.; Furey, T.; Rogic, S.; Kennedy, S.; Jones, S.; Rosenthal, A.; Wen, G.; Schilhabel, M.; Gloeckner, G.; Nyakatura, G.; Siebert, R.; Schlegelberger, B.; Korenberg, J.; Chen, X. N.; Fujiyama, A.; Hattori, M.; Toyoda, A.; Yada, T.; Park, H. S.; Sakaki, Y.; Shimizu, N.; Asakawa, S. *Nature* **2001**, *409*, 934−41.
(138) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.;

Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C. *Science* **2001**, *291*, 1304−51.

(139) Milgram, S. *Psychol. Today* **1967**, *2*, 60−67.

(140) Jeong, H.; Mason, S. P.; Barabasi, A. L.; Oltvai, Z. N. *Nature* **2001**, *411*, 41−2.

(141) Maslov, S.; Sneppen, K. *Science* **2002**, *296*, 910−3.

(142) Przulj, N.; Corneil, D. G.; Jurisica, I. *Bioinformatics* **2004**, *20*, 3508−15.

(143) Khanin, R.; Wit, E. *J. Comput. Biol.* **2006**, *13*, 810−8.

(144) Fraser, H. B.; Hirsh, A. E.; Steinmetz, L. M.; Scharfe, C.; Feldman, M. W. *Science* **2002**, *296*, 750−2.

(145) Wunderlich, Z.; Mirny, L. A. *Biophys. J.* **2006**, *91*, 2304−11.

(146) Cusick, M. E.; Klitgord, N.; Vidal, M.; Hill, D. E. *Hum. Mol. Genet.* **2005**, *14* (Spec No. 2), R171−81.

CR0682857