Shared components of protein complexes—versatile building blocks or biochemical artefacts?

Roland Krause,[1,2,4]* Christian von Mering,[2] Peer Bork,[2] and Thomas Dandekar[3]

Summary

Protein complexes perform many important functions in the cell. Large-scale studies of protein–protein interactions have not only revealed new complexes but have also placed many proteins into multiple complexes. Whilst the advocates of hypothesis-free research touted the discovery of these shared components as new links between diverse cellular processes, critical commentators denounced many of the findings as artefacts, thus questioning the usefulness of large-scale approaches. Here, we survey proteins known to be shared between complexes, as established in the literature, and compare them to shared components found in high-throughput screens. We discuss the various challenges to the identification and functional interpretation of bona fide shared components, namely contaminants, variant and megacomplexes, and transient interactions, and suggest that many of the novel shared components found in high-throughput screens are neither the results of contamination nor central components, but appear to be primarily regulatory links in cellular processes. ***BioEssays*** **26:1333–1343, 2004.** **© 2004 Wiley Periodicals, Inc.**

Introduction

Large-scale identification of protein complexes by mass spectrometry (MS) is a breakthrough in the exploration of protein–protein interactions and contributes substantially to our understanding of the molecular machinery. Several groups have published large-scale protein complex purifications (PCP) in the yeast *Saccharomyces cerevisiae*.[1–3] These interaction studies are significantly more informative than those employing yeast two-hybrid (Y2H) screening,[4–8] genetic interaction screening,[9] or in silico methods.[10–12] The advantages in coverage and accuracy was shown independently by several groups, particularly for the data obtained by Tandem Affinity Purification and mass spectrometric identification (TAP-MS)[13–15] The key concepts of protein–protein interactions are explained in the glossary (see below), and have been reviewed in great detail.[16–19] However, one particular issue, which is rarely addressed in writing but vigorously discussed in the field, is that many proteins are found in more than one complex in the PCP screens, for example, 37% of all proteins in the data set obtained by TAP-MS (see Figs. 1 and 2).[1–3] The authors concluded that these shared components would join the complexes into a higher order network, linking cellular processes. A more pessimistic interpretation was also voiced: the high frequency of shared proteins could simply reflect the artefactual identification of sub-stoichiometric contaminants in the biochemical purification.

We know that many processes in the cell are highly connected and, therefore, the finding of bridging components should not be surprising. A well-studied example is RNA polymerase II, which physically interacts with many different proteins performing diverse tasks such as histone acetylation, mRNA production, splicing and nuclear export.[20,21]

A thorough analysis of shared components is relevant for the functional interpretation of the data, for instance regarding the guilt-by-association principle:[22–24] i.e. does the shared protein perform several distinguishable functions, for example, an enzyme with an additional, non-catalytic role or as a common structural element used for scaffolding.[25] The notion of molecular function is undoubtedly vague and incomplete, and inferences of functional interaction between physically interacting proteins are not straightforward; however, true interactions within protein complexes are reliable measures of association and common function. Understanding shared components better should allow us to assign or even define multiple functions more objectively. On a more technical side, predicting protein complexes by bioinformatics methods[26–28] is complicated by the presence of shared components, as proteins potentially need to be assigned to more than a single complex, the exact number of complexes being unknown initially.

Studying shared components could also reveal candidate drug targets. Inhibition of such a protein would simultaneously act on more than one cellular process and the development of resistance against such an interference might be physiologi-

[1]Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
[2]EMBL, 69012 Heidelberg, Germany.
[3]Biozentrum, Am Hubland, 97074, Würzburg, Germany.
[4]Institute for Pure and Applied Mathematics, UCLA, Los Angeles, 90095-7121.
*Correspondence to: Roland Krause, EMBL, 69012 Heidelberg, Germany. E-mail: roland@krause.name
DOI 10.1002/bies.20141
Published online in Wiley InterScience (www.interscience.wiley.com).

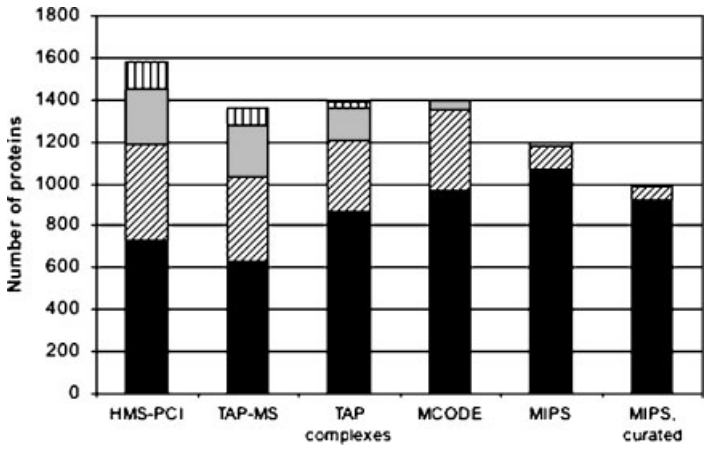

Figure 1. Number of shared protein components in collections of protein complexes. ■ Unique proteins. ▨ Proteins occurring 2 or 3 times. ▩ Proteins occurring 4 to 8 times. ▥ Proteins occurring more than 9 times. Experimental data were subject to filtering. Purifications retrieving only the tagged protein or homo-dimeric complexes were not considered.

cally constrained. The information supplied by shared components could thus be employed to detect candidate drug targets and assess the possibility of side effects.

In this work, we identify and discuss properties of proteins shared between distinct complexes. To this end, we need to examine the aspects of protein complexes in high-throughput studies that complicate the unambiguous detection of shared components. Contaminating background proteins occur in all high-throughput screens for interacting proteins and should generally be discussed in such high-level analyses of protein–protein interaction data. Likewise, the transient nature of protein interactions needs to be discussed for shared components. The arrangements of protein complexes into larger structures (megacomplexes) and the fact that some complexes are very similar and differ in only a few proteins (variant complexes) is an important consideration when discussing shared components also. After incorporating more detail into our view of the data sets, we can discuss features of shared proteins, and we conclude with the hypothesis that many are regulating, peripheral components of protein complexes. A glossary of some of the less-familiar technical terms is included at the end of the text.

Careful determination of contaminants

To obtain an initial overview of the known data, we counted the occurrence of proteins in more than one complex in data collections of protein complexes (Fig. 1). The established MIPS data set of protein complexes[29] is separate from high-throughput data, but contains some of the complexity of protein complexes, particular of variant complexes (see below). Therefore, the set was curated to contain only distinct complexes that do not overlap significantly, that is not by more than a Dice similarity of 0.8, a measure suited to score the similarity of protein complexes.[20,21] From this first overview of

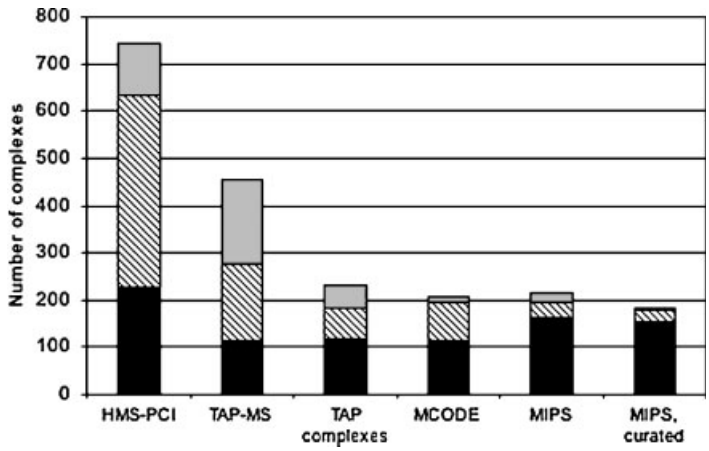

Figure 2. Size of overlap between complexes. ■ Unique complexes, not overlapping. ▧ Overlapping with other complexes by 1 to 3 components. ▩ Overlapping by more than 4 components, likely variant complexes. The high number of shared components in the experimental data sets is due to purification of the same complex in different experiments.

established complexes, we can set the lower limit for shared components of protein complexes to ~8% of the proteins within complexes.

There is a general difference between protein complexes established from the literature and the data from high-throughput experiments that arises primarily from the failure to detect certain components (false negatives) and the identifications of proteins that are not members of the complex (false positives). Tedious evaluation of each individual finding in a small-scale study achieves a confidence that cannot be reached in high-throughput studies.

For the analysis of shared components, missing identifications do not pose a problem, as they do not suggest additional shared components. However, false positives occur from systematic identifications of background proteins and might contribute massively to inflation in the frequency of shared components. Therefore, a close look at potential contaminants and their treatment in the analysis is paramount for the discussion of shared components, as all high-throughput screens contain false positives irrespective of their high value.[13,14] In the biochemical purification screens, potential contaminants were filtered by functional classes, their overall frequency of detection, and their occurrence in mock purifications of the respective protein tag. The false-positive list for the high-throughput PCP screens, which is independent of the scale and method used, comprises abundant enzymes and chaperones, as well as structural, proteasomal and, very prominently, ribosomal proteins.[1–3] Note that most background proteins do play important physiological roles and probably also associate in vivo, hence terms such as 'contaminant' or 'false positive' should be used with care. Surprisingly, the occurrence of proteins in the TAP-MS screen correlates only weakly with the overall protein expression levels in yeast[30] (Spearman rank correlation of 0.288). Yet, biochemical properties of particular groups of proteins contribute significantly to their repeated identifications as background proteins.[31]

We urge caution in defining contaminants globally, as a protein can be a bona fide interactor in the purification of one bait protein and a contaminant in the purification of another, even under the same protocol. An example is actin, a specific and stochiometric interactor in nuclear histone acetylase complexes,[32] yet a frequent contaminant in many other purifications, owing to its role as a major cytoskeletal component. Being part of several distinct histone acetylase complexes, actin is also a shared component in a strict sense (see Fig. 3).

The global filtering criterion also, unfortunately, removes other relevant information. The associations among glycolytic enzymes have been subject to an unsettled discussion on substrate channeling,[33] however, the filtering employed for PCP screens commonly removes these proteins. Likewise, ribosomal proteins can have extraribosomal functions[34] and the question as to whether to treat them as genuine interactors or as contaminants has to be answered independently for each study; see, for example, the ongoing discovery of complexes involved in the biogenesis of the ribosome.[1,2,35–37]

We inspected the most frequently shared components in the TAP-MS set.[2] This data set underwent careful curation previously but we suspect that several proteins, such as Adh1 or Yef3, should still be considered background proteins, as they join several unrelated complexes and share typical features with other contaminants (see above). Conversely, some proteins that were originally removed do not appear as contaminants on re-examination. Swd2 (YKL018w) was removed from the TAP-MS set because of its frequent detection, yet it is contained in the polyadenylation complex and the SET histone methylase complex only, which were repeatedly sampled in the screen.[2,38,39]

Another example consists of two related helicases, Rvb1 and Rvb2, which are frequently found in PCP screens. The proteins were previously described in the context of the INO80 histone acetylase complex.[40] Unfortunately, the parameter chosen led to the removal of Rvb1 but not Rvb2 from two screens[2,3]—a review of the data suggests that Rvb1 and Rvb2 should be considered as a functional module, present in INO80 and additional complexes.[41,42]

Despite potential removal of genuine interactors, limited filtering of contaminants is still required to extract reliable knowledge from the data sets. Rather than removing proteins globally, as has been the common practice, we suggest applying filtering after complex assembly or after delineation into weighted, binary interaction, preferably after integration with additional information such as abundance of the protein and its mRNA.[30,43] Filtering decreases the connectivity in the interaction networks resulting from Y2H and PCP screens, particularly for proteins with many interactors. It is suggested that these highly connected proteins (hubs) are of pivotal importance,[44,45] conveying robustness to the network.[46] Contaminants likewise appear central because of their high abundance, not because of bona fide interactions with functional implications. For a complete list of hubs, one would even have to connect each protein to the ribosome (where it is produced), most proteins to the proteasome (where they are degraded) and all nuclear proteins to the nuclear import machinery; however, these proteins are identical to typical contaminants. There is little functional insight to be gained from these truisms and removing these links deconvolutes the network without removing the proteins that we consider as true shared components.

Clearly, individual inspection suggests that most shared components do not appear as contaminants in the above fashion, and more-stringent filtering than is commonly used in PCP screens would only decrease the quality of the data. Most shared components in the TAP-MS screen occur in only a few complexes and are not widely spread. 64% of the shared

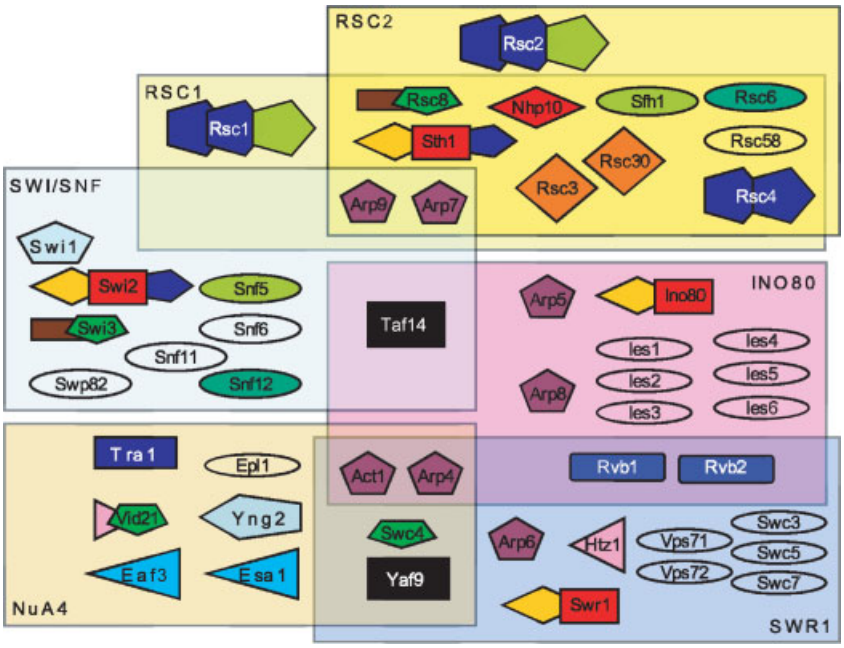

Figure 3. Shared components, domain conservation and variant complexes among chromatin remodelling complexes containing actin-related proteins (ARPs). Proteins with conserved domains are drawn similarly to the domain icons used in the SMART data base.[80] Proteins coloured as their complexes have no significant similarity to other proteins in the figure. The variant RSC complexes contain either Rsc1 or Rsc2. Despite their high sequence identity of 88.5%, the two proteins are functionally not fully equivalent.[81,82] The SWI/SNF complex is related to the RSC complexes, which is displayed in the domain structure of both complexes and their sharing of Arp7 and Arp9.[83] The sequence of Swp82 is not known, however, we suspect it to be encoded by YFL049w due to data in the TAP-MS screen.[2] Actin and the ARPs are typical features of these nuclear complexes. The newly discovered SWR1 complex[41,42] and the INO80 complex[40] contain additional ARPs. The NuA4 complex employs a different histone acetylase activity while preserving the elements such as the ARPs and proteins containing a SANT domain.[84] Several complexes (SAGA, SLIK, TFIIB, and TFIIF) share components with the complexes described here and are left out for clarity. They do not generally share other features such as ARPs with the displayed complexes.

components appear in two or three complexes only and, even if we were to consider all proteins occurring in four or more complexes as contaminants, we would still face 25% of all components as shared.

We conclude that the majority of the shared proteins in the PCP screens are correctly identified. Nevertheless, the mere detection of a protein in a complex does not necessarily imply that the protein is an integral part of the complex. Instead, the protein might be generally designed to bind other proteins, it might bind unspecifically under the experimental conditions, or its interaction is functionally relevant but temporally or spatially limited. Additional biochemical data can be used to distinguish between several types of transient interactions and potentially to reveal whether a shared protein interacts transiently with the complexes that it is detected with or whether it is an integral part of them.

Catching transient interactions

Some recent interpretations of large-scale interaction data overemphasize the static aspects of protein complexes and tend to ignore the fact that the dynamic range of expression as well as the association rates of interactions span several orders of magnitude.[30,47,48] Even for text-book examples of molecular machines,[49] individual proteins associate and separate repeatedly in their lifetime, as demonstrated impressively for core RNA polymerase I (RNAP I) with GFP-fusion proteins and in vivo microscopy.[50]

It is often argued that the main discrepancies between protein interaction techniques such PCP and Y2H are due to their suitability (or lack thereof) of detecting transient interactions. However, there is more than one definition for unstable contacts between proteins: constant maintenance throughout the cell cycle can be used to categorize transient and stable complexes by analyzing the mRNA expression of synchronized cells.[51] In addition, interactions have been classified as transient if at least one of the proteins performs its function independently, so multi-subunit enzymes exemplify stable complexes in this scenario.[31] Kinetic data (e.g. association/dissociation constants), the most objective classification, are available for only a limited number of complexes,[52] mostly for

receptor–ligand binding, which are difficult to study using large-scale exploratory protein–protein interaction screens by PCP.

Each definition captures an important aspect of flexibility in protein–protein interactions. It would be desirable to agree on clear naming conventions rather than using the general term "transient interactions"—often to explain the disappointing absence of expected interactors or to downplay apparent differences between data sets. From previous analyses, we expect a high number of transient interactions by all the definitions given above—many established complexes could be considered transient, even if they can be readily retrieved by affinity purification.[51,53,54] Therefore, the approach of simply considering the interactions within the MIPS set of protein complexes as stable (after minor curation), and interactions obtained by other sources as transient might be somewhat ad hoc.[55]

For PCP data, one could consider proteins as transient interactors if they are found in several complexes and are known to be involved in many cellular processes. Examples include Cdc48, a protein that seems to connect several cellular processes to the cell cycle, and Srv2, which is known to be associated with adenylate cyclase and could mediate signaling—the association with these protein complexes was not known previously.[2] However, the number of shared components that can be explained by such arrangements is limited and other phenomena contribute to the complexity observed. We believe that, in particular, large overlaps between related protein complexes contribute to the occurrence of shared components.

Unravelling megacomplexes and variant complexes

Protein complexes interact with themselves to form larger structures, dubbed megacomplexes.[56] The two-subunit ribosome is simple to delineate, yet the individual subunits interact or merge with several complexes during their biogenesis.[35,36] Counting shared components between a megacomplex and a subcomplex without prior consideration of their arrangement, results in a large number of shared proteins being detected. This suggests that complexes that are simply assemblies of smaller ones should be eliminated in comparisons. To deal with megacomplexes, databases such as MIPS[29] and the Gene Ontology database[59] use a hierarchical approach.

Another simple solution to the problem would be to map all complexes that have functional and physical contacts into one megacomplex. With this model, we would consider the RNA polymerase II holo-enzyme complex as a superset of all proteins involved in mRNA transcription, ranging from histone modification to nuclear export,[20] a complex that contains several hundred proteins.[21] However, as the resolution of the methods at our disposal allows for more precision, we consider it to be more useful to define models for protein complexes close to the experimental setup used to detect them. This is certainly a pragmatic definition that requires many compromises.

Another important complication when assessing shared components arises from the existence of variant complexes, which differ only slightly in their composition (see Table 2; Fig. 3). When simply counting shared components between variant complexes, the number of shared proteins will be very high, but functional differences are often not easily detectable. This aspect is already well studied in yeast but needs to be considered especially when higher eukaryotes are employed. Similar to the expansion of protein domain arrangements in metazoans, their complexes undergo multifarious, semi-redundant arrangements, exemplified by the NF-κB complex, consisting of five proteins that arrange into functionally distinct homodimers and heterodimers.[58]

Variant complexes can be identified by homologies between the subunits of the different complexes. Within the 48 complexes having shared components in the curated MIPS set, 15 show clear homologies to other complexes in the set, and many more show at least conservation of protein domains. For functional analyses of shared components, we suggest considering variant protein complexes as "functionally equivalent"—these complexes should be considered as a single entity, when used in benchmarks.

The complexes obtained by TAP-MS do not appear overly enriched in variant complexes, as the number of shared components between the complexes is usually smaller than the number of unique proteins. However, the proteasome, the spliceosome and the protein complexes involved in the biogenesis of the ribosome are clearly recognizable as variant complexes, and consequently their components appear to be shared.

Genuine shared components may act as connectors

Having addressed the major complicating factors for shared components—contaminants, transient interactions and the arrangement of the complexes—they can be identified more successfully and common properties derived. The structural arrangement of shared components is important in identifying the function of shared components (see Fig. 4A). For instance, Swd2, is an element of two complexes but probably not their binding interface,[38,39] whereas e.g. Sus1 is thought to tether two complexes together physically.[59] Unfortunately, neither Y2H nor PCP reveal the local arrangement and it is not always possible to predict the three-dimensional structure of the complex using the structures of the individual proteins.[60,61]

What are the possible evolutionary scenarios for the emergence of shared components (see Fig. 4B)? Duplication of one or several genes within a proto-complex is the most-straight-forward explanation and has been used as a model for

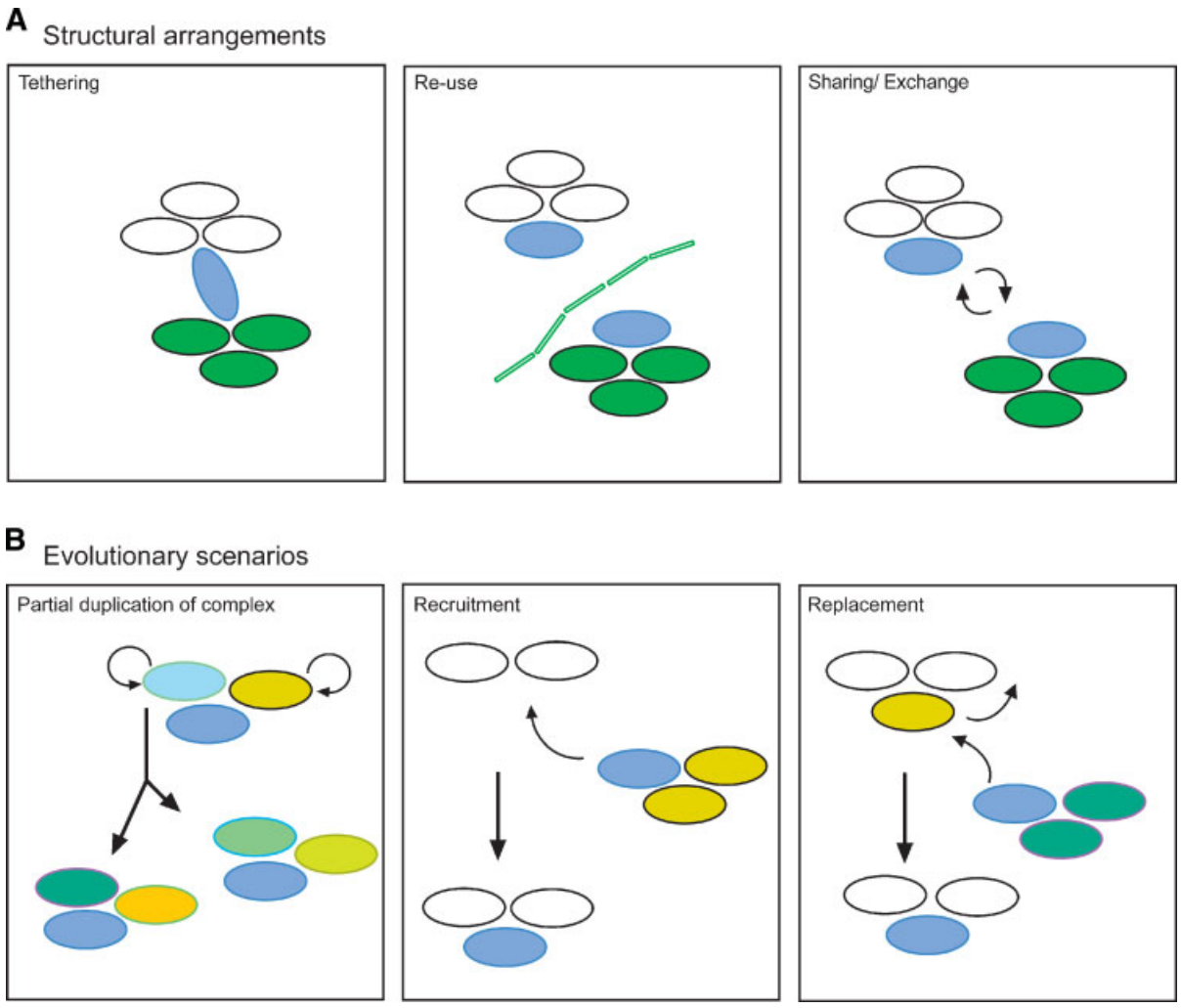

Figure 4. A: Possible structural arrangements. Tethering: The two complexes share a common polypeptide unit, functional interaction between the complexes. Re-use: Independent sharing between different complexes, functional separation. Sharing/Exchange: The shared protein can be modified for information transfer. **B:** Evolutionary scenarios for emergence of shared components. Partial duplication of protein components: One or several but not all components undergo gene duplication, leading to distinct complexes that share components. The duplication event can be traced if the proteins involved show sequence or structural similarity. Recruitment: An unrelated protein is recruited into a complex due to change of binding properties of the participating proteins. If the protein is already part of a complex, it appears as shared but there is no evolutionary relation between the proteins of the two complexes. Replacement: Changes in the binding properties of protein could lead to the prevalence of a protein binding at a specific position replacing a protein in the same position. Such a scenario would be difficult to demonstrate.

the emergence of binary interactions.[62] However, even non-homologous complexes share proteins, and these might perform a common function, such as DNA binding. We propose that independent proteins are recruited into an existing complex due to emergence of suitable binding interfaces. This recruitment provides an alternative evolutionary origin and can be seen as an inventive step, similar to gene invention (e.g. genes arising from non-coding elements rather than duplication). As several proteins appear to be recruited, it would be interesting to compare the ratio of recruitment and duplication in protein complexes to the ratio of gene duplication and gene invention, which is very large.[63] An explanation can be given by the natural tendencies of proteins to aggregate.[64]

A third possibility would be replacement of components; however, such a scenario will be difficult to quantify against the two other options. Unfortunately, yeast is the only organism where protein complexes have been established on a large scale, thus our ability to understand their evolution is limited, as we can compare different organisms only in the profiles of conserved proteins, not in their patterns of

Table 1. Examples illustrating properties of established shared components

Shared proteins	Complexes	Structural arrangement	Comments	Sample references
Sus1	SAGA histone acetylase Nuclear export	Tethering	Suggested to physically connect export and early gene expression	(59)
Med8	Hexokinase 20S Proteasome Mediator (RGR1 module)	Re-use	Part of the large Mediator complex and in a distinct regulatory complex with hexokinase (Hxk2), might also interact with the Proteasome	(85–87)
Pfk1	Phosphofructokinase (with Pfk2) Tubulin	Re-use	Pfk1 is reported to form a defined complex with Tubulin.	(2,88–90)
Taf14	SWI-SNF, INO80 TFIIF SAGA	Sharing/exchange	The protein is apparently not tethering the complexes together but occurs in separate entities.	(2,76,83)
Rpn5	COP9/Signalosome (CSN) 19S regulatory subunit part of the Proteasome	Sharing/exchange	CSN and 19S are distantly homologous structures.	(2,91,92)
Shared proteins	**Complexes**	**Evolutionary model**	**Comments**	**Sample references**
Rpo26, Rpb5, Rpb10, Rpb8, Rpc10	RNAP I RNAP II RNAP III	Duplication	Common for all core RNAPs, homologies between other members of the complexes. The single RNA polymerase in prokaryotes suggests evolution by duplication.	(93)
Swd2	Polyadenylation, Set1 complex (COMPASS)	Recruitment	No significant homologies between members of the two complexes.	(38,39)
Act1	SWR1 NuA4 INO80	Recruitment and duplication	The additional role of actin as part of the histone acetylases can be explained best by a recruitment, actin-related proteins in different complexes probably arose via gene duplication.	(2,32,41,42)
Shared proteins	**Complexes**	**Functional implication**	**Comments**	**Sample references**
Arp7, Arp9	Abundant chromatin remodelling complex (RSC) SWI/SNF chromatin remodelling complex	Binding to nucleosomes, regulation of complex activity	Regulatory function has been proposed for actin related proteins with chromatin-remodelling complexes.	(70,83,94)
Ldp1	Pyruvate dehydrogenase 2-oxoglutarate dehydrogenase Glycine decarboxylase Branched chain oxo-acid dehydrogenase	Common enzymatic function	Lipoamid dehydrogenase performs the same biochemical function in all complexes.	(67,95)
Rpl30	Ribosome U1 splicesomal particle	Regulation	Regulates splicing of the RPL30 transcript	(96)

The first row set focuses on proteins and complexes with insight into the structural arrangement (see also Fig. 3A). The second set of rows contains examples of proteins and complexes where suggestions for their evolution can be made (see also Fig. 3B). The last set lists examples of functional roles of shared components.

conserved interactions. Nevertheless, such studies have been attempted and the analyses gave contradictory findings.[45,65,66]

One could expect evolutionarily conserved, shared components to participate as central elements of protein complexes (structurally or functionally). Lipoamid dehydrogenase (Lpd1) is an example of a shared protein that acts as a central building block in several multi-protein enzymes[67]—can we generalize this finding? RNAP I, RNAP II, and RNAP III provide instructive examples as they share five components and their three-dimensional structure is well resolved.[68] Studying them does not support the initial assumption. The central structural and major functional elements are unique to the polymerases and the shared components associate peripherally, suggesting a connecting or regulatory role, which has been confirmed for Rpb5.[69] Other complexes with shared components display a similar setup, such as the RSC and the SWI/SNF complex[70,71] (see also Fig. 3 and Table 1), and we would speculate that shared components are more often employed in regulatory or auxiliary functions than as common structural elements. They may represent messengers between processes and, because they employ the same protein in several processes, they could be advantageous to the cell in providing a single point at which to regulate them.

Another related role for shared components is probably in physically bridging protein complex interactions, such as for Spt6, which directly tethers the nuclear exosome to RNAP II.[72]

When attempting to quantify these functional aspects, we discovered complications due to the current annotation of proteins and their complexes. Often, all proteins in a given complex with enzymatic activity are considered to be enzymes even if only one protein contains the active site and displays activity if isolated (such as PP2A, see Table 2). Currently, only meticulous, manual analysis of the protein complexes following the steps outlined above leads to the detection of novel bona fide shared components. Integrating all data in the light of protein complexes with shared components and studying their behavior is a challenging frontier to bioinformatics.

New techniques for the discovery of protein–protein interactions are becoming available and we expect to observe sharing of components in compartments that are not well sampled yet, such as membranes[73] or in multicellular organisms.[74]

Conclusions

How many shared components are we to expect? Studying shared components is connected to the challenges of interpreting protein interaction data because megacomplexes, variant complexes and transient interactions have to be considered. Consequently the estimates vary drastically, and yeast is the only organism with sufficient data for such estimates. The data model chosen for the complexes is pivotal, as different treatments of megacomplexes and variant complexes would mean that studies could not be compared.

The average number of interactions for a protein ranges from 1.6.[6] to 12,[2] the total number may be within a range of 10.000[75] to 30.000[14] in *Saccharomyces cerevisiae*, whereas the number of corresponding complexes is likely to be between 173, in the curated MIPS set, and about 500, when combining the HMS-PCI and the TAP-MS screen.[28] Given the basic problems discussed here, we will need to consolidate the available data before we can quantify features such as shared components in a more accurate way. However, we can conclude that shared proteins are commonplace, and we bracket their occurrence between 8% and 30% of interacting proteins on the basis of the large-scale studies (see Fig. 2).

How many distinct complexes share a single component? All of the components that can be described as uniquely shared, participate in no more than five non-variant complexes (with Taf14 being the most widely shared),[76] typically such proteins contribute to two or three complexes only. Most of the shared components between sufficiently different complexes associate peripherally and are not integral member of the

Table 2. Variant complexes in yeast

Complex	Variant proteins	Reference
RSC (RSC1 and RSC2)	Rsc1/Rsc2	(81,82)
Lsm2p-Lsm8p, Lsm1p-Lsm7p	Lsm1/Lsm8	(97)
Trehalose synthase	Tps3/Tsl1	(98)
Protein phosphatase 2A	Cdc55, Rts3/ Pph21, Pph22	(99)
SAGA/SLIK	Spt8/Spt7	(100)

Complexes share the majority of their components except for the proteins in the middle columns. All variant proteins are homologous. Note that the distinction between the variant complexes is not always addressed in the literature. For instance, protein phosphatase 2A consists of four variant complexes with combinations of either Cdc55 or Rts3 with either Pph21 or Pph22, which have been shown to be functionally separable. However, the complex is referred to as protein phosphatase 2A.

complexes. The special arrangement of these proteins reveals new insights into our understanding of regulation in the cellular machinery.

Glossary

Dice similarity

A similarity measure that can be used to compare protein complexes. The Dice similarity m_d is defined as $m_d = \frac{2n_i}{n_a+n_b}$ with n_a and n_b being the number of proteins in the individual protein complexes and n_i the number of protein in the intersection of the two groups. Depending on the purpose of the comparison, other similarity measures might be useful.[28]

HMS-PCI—High-throughput mass spectrometric protein complex identification

A PCP technique that adds a short tag (called FLAG) to a plasmid-borne protein which is overexpressed. It was used in the first large-scale PCP study,[3] (together with the TAP-MS study). The data is available from http://www.mdsp.com/yeast/

MCODE

An algorithm that finds protein complexes in interaction networks. A global prediction of all complexes in yeast given all interaction data available in early 2002 is used here as reference data set.[26]

MIPS data set

The manually assembled data set of protein complexes has become the gold standard set for bioinformatics research, as it does not contain information from high-throughput screens but covers the known literature well.[29] It can be found at http://mips.gsf.de/desc/yeast/

MS—Mass spectrometry

The method of choice to identify proteins in large-scale proteomics projects. The molecular weight of fragments of a protein is accurately determined, allowing identification of the protein in a sequence database.[77]

PCP—Protein Complex Purification

An umbrella term for techniques for the elucidation of protein complexes by isolation of the complex, typically involving targeting one of the proteins and identification of the interactors. The individual experiments are often referred to as ''purifications''.[18]

TAP—Tandem Affinity Purification

A complex purification technique, which uses homologous recombination for introduction of a protein tag and a gentle, two-step procedure to retrieve interacting partner.[78] A study using TAP and MS identification (TAP-MS) provides an important data set.[2] The data is available at http://yeast.cellzome.com

Y2H—Yeast two hybrid

A method for the study of protein–protein interactions that is used for screening. Uses two chimeric proteins (hybrids) that are joined with an activation domain and a DNA binding, respectively. The interaction of the two protein triggers expression of a reporter gene.[79]

References

1. Krogan NJ, Peng WT, Cagney G, Robinson MD, Haw R, et al. 2004. High-definition macromolecular composition of yeast RNA-processing complexes. Mol Cell 13:225–239.
2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147.
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183.
4. Ito T, Ota K, Kubota H, Yamaguchi Y, Chiba T, et al. 2002. Roles for the two-hybrid system in exploration of the yeast protein interactome. Mol Cell Proteomics 1:561–566.
5. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627.
6. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574.
7. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. 2003. A protein interaction map of *Drosophila melanogaster*. Science 302: 1727–1736.
8. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. 2004. A map of the interactome network of the metazoan C. elegans. Science 303:540–543.
9. Tong AH, Lesage G, Bader GD, Ding H, Xu H, et al. 2004. Global mapping of the yeast genetic interaction network. Science 303:808–813.
10. Huynen MA, Snel B, von Mering C, Bork P. 2003. Function prediction and protein networks. Curr Opin Cell Biol 15:191–198.
11. Osterman A, Overbeek R. 2003. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 7:238–251.
12. Marcotte EM. 2000. Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol 10:359–365.
13. Bader GD, Hogue CW. 2002. Analyzing yeast protein–protein interaction data obtained from different sources. Nat Biotechnol 20:991–997.
14. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. 2002. Comparative assessment of large-scale data sets of protein protein interactions. Nature 417:399–403.
15. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, et al. 2002. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18:529–536.
16. Bork P, Jensen LJ, Von Mering C, Ramani AK, Lee I, et al. 2004. Protein interaction networks from yeast to human. Curr Opin Struct Biol 14: 292–299.
17. Alm E, Arkin AP. 2003. Biological networks. Curr Opin Struct Biol 13: 193–202.
18. Janin J, Seraphin B. 2003. Genome-wide studies of protein–protein interaction. Curr Opin Struct Biol 13:383–388.
19. Gavin AC, Superti-Furga G. 2003. Protein complexes and proteome organization from yeast to man. Curr Opin Chem Biol 7:21–27.
20. Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. Nature 416:499–506.
21. Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, et al. 2003. The study of macromolecular complexes by quantitative proteomics. Nat Genet 33:349–355.

22. Deng M, Sun F, Chen T. 2003. Assessment of the reliability of protein–protein interactions and protein function prediction. Pac Symp Biocomput 8:140–151.
23. Vazquez A, Flammini A, Maritan A, Vespignani A. 2003. Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 21:697–700.
24. Samanta MP, Liang S. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci USA 100:12579–12583.
25. Jeffery CJ. 2003. Moonlighting proteins: old proteins learning new tricks. Trends Genet 19:415–417.
26. Bader GD, Hogue CW. 2003. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4:2.
27. Spirin V, Mirny LA. 2003. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100:12123–12128.
28. Krause R, von Mering C, Bork P. 2003. A comprehensive set of protein complexes in yeast: mining large scale protein–protein interaction screens. Bioinformatics 19:1901–1908.
29. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, et al. 2002. MIPS: a database for genomes and protein sequences. Nucleic Acids Res 30:31–34.
30. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. 2003. Global analysis of protein expression in yeast. Nature 425:737–741.
31. Aloy P, Russell RB. 2002. The third dimension for protein interactions and complexes. Trends Biochem Sci 27:633–638.
32. Olave IA, Reck-Peterson SL, Crabtree GR. 2002. Nuclear actin and actin-related proteins in chromatin remodeling. Annu Rev Biochem 71: 755–781.
33. Srere PA, Ovadi J. 1990. Enzyme–enzyme interactions and their metabolic role. FEBS Lett 268:360–364.
34. Wool IG. 1996. Extraribosomal functions of ribosomal proteins. Trends Biochem Sci 21:164–165.
35. Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, et al. 2002. A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. Nature 417:967–970.
36. Schafer T, Strauss D, Petfalski E, Tollervey D, Hurt E. 2003. The path from nucleolar 90S to cytoplasmic 40S pre-ribosomes. EMBO J 22: 1370–1380.
37. Grandi P, Rybin V, Bassler J, Petfalski E, Strauss D, et al. 2002. 90S pre-ribosomes include the 35S pre-rRNA, the U3 snoRNP, and 40S subunit processing factors but predominantly lack 60S synthesis factors. Mol Cell 10:105–115.
38. Roguev A, Schaft D, Shevchenko A, Pijnappel WW, Wilm M, et al. 2001. The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. EMBO J 20:7137–7148.
39. Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, et al. 2001. COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. Proc Natl Acad Sci USA 98:12902–12907.
40. Shen X, Mizuguchi G, Hamiche A, Wu C. 2000. A chromatin remodelling complex involved in transcription and DNA processing. Nature 406:541–544.
41. Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, et al. 2004. A Protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. PLoS Biol 2:E131.
42. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, et al. 2004. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science 303:343–348.
43. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728.
44. Maslov S, Sneppen K. 2002. Specificity and stability in topology of protein networks. Science 296:910–913.
45. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.
46. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.
47. Schreiber G. 2002. Kinetic studies of protein–protein interactions. Curr Opin Struct Biol 12:41–47.
48. Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720–1730.
49. Alberts B. 1998. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294.
50. Dundr M, Hoffmann-Rohrer U, Hu Q, Grummt I, Rothblum LI, et al. 2002. A kinetic framework for a mammalian RNA polymerase in vivo. Science 298:1623–1626.
51. Jansen R, Greenbaum D, Gerstein M. 2002. Relating whole-genome expression data with protein–protein interactions. Genome Res 12: 37–46.
52. Nooren IM, Thornton JM. 2003. Structural characterisation and functional significance of transient protein–protein interactions. J Mol Biol 325:991–1018.
53. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, et al. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol Cell 9:1133–1143.
54. Ge H, Liu Z, Church GM, Vidal M. 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat Genet 29:482–486.
55. Teichmann SA. 2002. The constraints protein–protein interactions place on sequence divergence. J Mol Biol 324:399–407.
56. Deshaies RJ, Seol JH, McDonald WH, Cope G, Lyapina S, et al. 2002. Charting the protein complexome in yeast by mass spectrometry. Mol Cell Proteomics 1:3–10.
57. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.
58. Saccani S, Pantano S, Natoli G. 2003. Modulation of NF-kappaB activity by exchange of dimers. Mol Cell 11:1563–1574.
59. Rodriguez-Navarro S, Fischer T, Luo MJ, Antunez O, Brettschneider S, et al. 2004. Sus1, a Functional component of the SAGA histone acetylase complex and the nuclear pore-associated mRNA export machinery. Cell 116:75–86.
60. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, et al. 2004. Structure-based assembly of protein complexes in yeast. Science 303:2026–2029.
61. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, et al. 2003. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52:2–9.
62. Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292.
63. Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. Science 300:1701–1703.
64. Dobson CM. 2004. Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol 15:3–16.
65. Jordan IK, Wolf YI, Koonin EV. 2003. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1.
66. Pagel P, Mewes HW, Frishman D. 2004. Conservation of protein–protein interactions — lessons from ascomycota. Trends Genet 20: 72–76.
67. Sinclair DA, Dawes IW. 1995. Genetics of the synthesis of serine from glycine and the utilization of glycine as sole nitrogen source by *Saccharomyces cerevisiae*. Genetics 140:1213–1222.
68. Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, et al. 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. Science 288:640–649.
69. Dorjsuren D, Lin Y, Wei W, Yamashita T, Nomura T, et al. 1998. RMP, a novel RNA polymerase II subunit 5-interacting protein, counteracts transactivation by hepatitis B virus X protein. Mol Cell Biol 18:7546–7555.
70. Boyer LA, Peterson CL. 2000. Actin-related proteins (Arps): conformational switches for chromatin-remodeling machines? Bioessays 22: 666–672.
71. Szerlong H, Saha A, Cairns BR. 2003. The nuclear actin-related proteins Arp7 and Arp9: a dimeric module that cooperates with architectural proteins for chromatin remodeling. EMBO J 22:3175–3187.

72. Lindstrom DL, Squazzo SL, Muster N, Burckin TA, Wachter KC, et al. 2003. Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. Mol Cell Biol 23:1368–1378.
73. Nohe A, Petersen NO. 2004. Analyzing for co-localization of proteins at a cell membrane. Curr Pharm Biotechnol 5:213–220.
74. Remy I, Michnick SW. 2004. Mapping biochemical networks with protein-fragment complementation assays. Methods Mol Biol 261:411–426.
75. Sprinzak E, Sattath S, Margalit H. 2003. How reliable are experimental protein–protein interaction data? J Mol Biol 327:919–923.
76. Henry NL, Campbell AM, Feaver WJ, Poon D, Weil PA, et al. 1994. TFIIF-TAF-RNA polymerase II connection. Genes Dev 8:2868–2878.
77. Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. Nature 422:198–207.
78. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, et al. 1999. A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17:1030–1032.
79. Fields S, Song O. 1989. A novel genetic system to detect protein–protein interactions. Nature 340:245–246.
80. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, et al. 2004. SMART 4.0: towards genomic data integration. Nucleic Acids Res 32 Database issue:D142–144.
81. Yukawa M, Koyama H, Miyahara K, Tsuchiya E. 2002. Functional differences between RSC1 and RSC2, components of a for growth essential chromatin-remodeling complex of *Saccharomyces cerevisiae*, during the sporulation process. FEMS Yeast Res 2:87–91.
82. Cairns BR, Schlichter A, Erdjument-Bromage H, Tempst P, Kornberg RD, et al. 1999. Two functionally distinct forms of the RSC nucleosome-remodeling complex, containing essential AT hook, BAH, and bromodomains. Mol Cell 4:715–723.
83. Smith CL, Horowitz-Scherer R, Flanagan JF, Woodcock CL, Peterson CL. 2003. Structural analysis of the yeast SWI/SNF chromatin remodeling complex. Nat Struct Biol 10:141–145.
84. Eisen A, Utley RT, Nourani A, Allard S, Schmidt P, et al. 2001. The yeast NuA4 and Drosophila MSL complexes contain homologous subunits important for transcription regulation. J Biol Chem 276:3484–3491.
85. Dotson MR, Yuan CX, Roeder RG, Myers LC, Gustafsson CM, et al. 2000. Structural organization of yeast and mammalian mediator complexes. Proc Natl Acad Sci USA 97:14307–14310.
86. de la Cera T, Herrero P, Moreno-Herrero F, Chaves RS, Moreno F. 2002. Mediator factor Med8p interacts with the hexokinase 2: implication in the glucose signalling pathway of *Saccharomyces cerevisiae*. J Mol Biol 319:703–714.
87. Verma R, Chen S, Feldman R, Schieltz D, Yates J, et al. 2000. Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. Mol Biol Cell 11:3425–3439.
88. Klinder A, Kirchberger J, Edelmann A, Kopperschlager G. 1998. Assembly of phosphofructokinase-1 from *Saccharomyces cerevisiae* in extracts of single-deletion mutants. Yeast 14:323–334.
89. Lehotzky A, Telegdi M, Liliom K, Ovadi J. 1993. Interaction of phosphofructokinase with tubulin and microtubules. Quantitative evaluation of the mutual effects. J Biol Chem 268:10888–10894.
90. Vertessy BG, Kovacs J, Low P, Lehotzky A, Molnar A, et al. 1997. Characterization of microtubule-phosphofructokinase complex: specific effects of MgATP and vinblastine. Biochemistry 36:2051–2062.
91. Wee S, Hetfeld B, Dubiel W, Wolf DA. 2002. Conservation of the COP9/signalosome in budding yeast. BMC Genet 3:15.
92. Fu H, Reis N, Lee Y, Glickman MH, Vierstra RD. 2001. Subunit interaction maps for the regulatory particle of the 26S proteasome and the COP9 signalosome. EMBO J 20:7096–7107.
93. Cramer P, Bushnell DA, Kornberg RD. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science 292:1863–1876.
94. Cairns BR, Erdjument-Bromage H, Tempst P, Winston F, Kornberg RD. 1998. Two actin-related proteins are shared functional components of the chromatin-remodeling complexes RSC and SWI/SNF. Mol Cell 2: 639–651.
95. Zaman Z, Bowman SB, Kornfeld GD, Brown AJ, Dawes IW. 1999. Transcription factor GCN4 for control of amino acid biosynthesis also regulates the expression of the gene for lipoamide dehydrogenase. Biochem J 340(Pt 3):855–862.
96. Vilardell J, Chartrand P, Singer RH, Warner JR. 2000. The odyssey of a regulated transcript. Rna 6:1773–1780.
97. Bouveret E, Rigaut G, Shevchenko A, Wilm M, Seraphin B. 2000. A Sm-like protein complex that participates in mRNA degradation. EMBO J 19:1661–1671.
98. Bell W, Sun W, Hohmann S, Wera S, Reinders A, et al. 1998. Composition and functional analysis of the *Saccharomyces cerevisiae* trehalose synthase complex. J Biol Chem 273:33311–33319.
99. Gentry MS, Hallberg RL. 2002. Localization of *Saccharomyces cerevisiae* protein phosphatase 2A subunits throughout mitotic cell cycle. Mol Biol Cell 13:3477–3492.
100. Pray-Grant MG, Schieltz D, McMahon SJ, Wood JM, Kennedy EL, et al. 2002. The novel SLIK histone acetyltransferase complex functions in the yeast retrograde response pathway. Mol Cell Biol 22:8774–8786.