Minireview

The elusive yeast interactome

Johannes Goll* and Peter Uetz*†

Addresses: *Institut für Genetik, Forschungszentrum Karlsruhe, Box 3640, 76021 Karlsruhe, Germany. †The Institute of Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

Correspondence: Peter Uetz. Email: peter@uetz.de

Published: 30 June 2006

Genome ***Biology*** 2006, **7:**223 (doi:10.1186/gb-2006-7-6-223)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/6/223

Abstract

Simple eukaryotic cells such as yeast could contain around 800 protein complexes, as two new comprehensive studies show. But slightly different approaches resulted in surprising differences between the two datasets, showing that more work is required to get a complete picture of the yeast interactome.

Protein complexes are the workhorses of the cell as they are involved in almost all biological processes from transmembrane signaling to gene expression. Only a few are really well understood in terms of structure and function, however, and many appear to be involved in processes we do not know much about. In two independent recent papers [1,2], groups from the European Molecular Biology Laboratory (EMBL), Cellzome (a spin-off company from EMBL), and the University of Toronto have published comprehensive surveys of all the protein complexes detected in yeast - the yeast interactome or 'complexome' as one might now call it (Table 1; see also [3,4]). This is a landmark achievement, given that no other cell or organism has been surveyed at such a level of detail. More important, yeast is a prototypic eukaryotic cell that is a model for human cells, and most yeast complexes probably have homologs in humans.

From proteome to complexome

The characterization of protein complexes sounds trivial: insert a piece of DNA encoding a 'tag' into a protein-coding gene and let the cells express the tagged protein (the 'bait'). Then break up the cell and 'pull out' the tagged protein with all its associated proteins (the 'preys') by some technique such as co-immunoprecipitation or tandem affinity purification (TAP). Finally, identify all the proteins in the purified complex by mass spectrometry [5]. Then repeat this procedure for all the protein-coding genes in the yeast genome. This is exactly what the two teams did [1,2]. But although the identification of protein complexes sounds easy, it is not. Complications arise, for example, when proteins belonging to the same complex are tagged and the resulting complexes are purified. In most cases this leads to conflicting information, because these purifications have slightly different protein compositions, depending on which protein was the tagged one (Figure 1 and Table 2). Different 'complexes' are recovered even when the same tagged protein is purified repeatedly. For example, Gavin *et al.* [1] repeated 139 of their purifications (99 with soluble and 40 with membrane proteins), and as foreshadowed in their previous pilot study [6] only 69% of the recovered proteins were common to both purifications. The pull-down approach is thus fairly reproducible but does have a significant error margin. In addition, many proteins are part of several different complexes: one bait protein may thus pull down several independent complexes that appear in the experiment to be one large complex.

Although the strategy is similar, there are a number of differences between the approaches taken by Gavin *et al.* [1] and Krogan *et al.* [2]. First, the protocols were not identical. Second, only Gavin *et al.* attempted to tag all transmembrane proteins. Third, Gavin *et al.* provide raw purification data whereas Krogan *et al.* provide only computationally processed information at the time of writing: for example, the latter removed 44 preys detected in more than 3% of purifications and nearly all ribosomal proteins. These

Table 1

Comparison of two projects aimed at determining the number of protein complexes in *S. cerevisiae*

	Gavin *et al.* [1]	Krogan *et al.* [2]
Number of proteins TAP-targeted	6,466	?
TAP fusion expression/purification attempts	3,206	4,562
Successful purifications	?	2,357*
Proteins with more than zero partners	1,993	?
Distinct proteins identified	2,760	4,087 (2,708 in core set)
'Distinct' complexes identified†	491	547
Average number of proteins per complex	3.1 (core proteins)	4.9

*At least one protein was identified in each of these purifications; 1,613 baits were successfully analyzed using SDS-PAGE and matrix-assisted laser desorption ionization mass spectrometry (MALDI) and 2,001 baits by liquid chromatography followed by tandem mass spectrometry (LC-MS/MS); 1,257 purifications were successful using both, 356 using only MALDI and 744 using only LC-MS/MS. †Note that these are not necessarily distinct physical complexes but computationally derived complexes. A question mark indicates that a number for this was not given in the article.

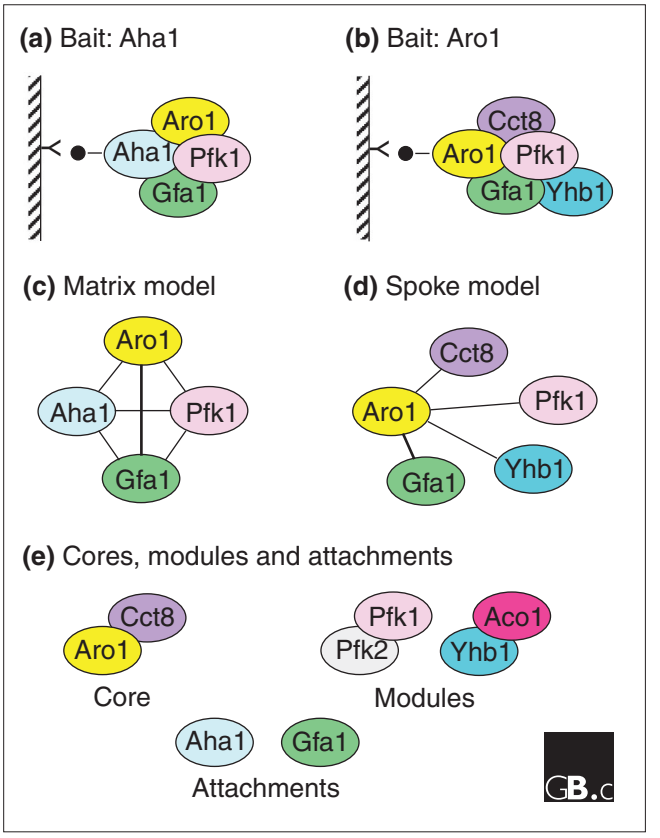

Figure 1
The difference between complex purifications and protein complexes. **(a,b)** When two proteins belonging to the same complex are tagged in independent yeast strains and the other components of the complex identified, the two purifications rarely return precisely the same list of components. **(c,d)** Although proteins in a complex are associated, it is usually unclear which proteins interact directly with each other. To predict direct interactions, either the matrix (c) or spoke model (d) is applied to lists of co-purified proteins. To evaluate such interactions Gavin *et al.* [1] invented the socio-affinity index (SAI). In brief, the SAI quantifies the tendency for a protein pair (for example, Aro1 and Gfa1) to identify each other when one of them is tagged (as in b) and to co-purify when other proteins are tagged (as in a) relative to what would be expected from their frequency in the dataset: that is, how many times this protein was found as prey. High-affinity SAI values result when both proteins co-purify when either one is tagged (without co-purifying many other proteins) and when both are always seen together in purifications made with other baits. **(e)** For this particular complex one core, two modules and two attachments have been identified. Note that modules cannot be computed from only two purifications; the assemblies of Figure 1e are derived from the eight purifications shown in Table 2 and additional purifications not shown.

proteins were considered as nonspecific contaminants and thus as false positives. In contrast, such nonspecific contaminants were left in the raw dataset of Gavin *et al.* and only later removed (or not) when they determined their final list of complexes (see Figure 1 and below). Both groups aimed at the same goal: to unravel all the protein complexes in yeast. Using similar technology they should have got the same results, despite certain differences in method. But did they? As we shall see, not quite.

To distill defined complexes from their raw purification data, both first transformed their raw data into weighted binary interactions. While Krogan *et al.* [2] used a machine learning algorithm trained by hand-curated protein complexes, Gavin *et al.* [1] invented a new measure, solely based on raw purification data, which they called the 'socio-affinity index'. In the next step, cluster algorithms were used to determine distinct complexes. Using an iterative clustering procedure, Gavin *et al.* [1] came up with the classification outlined in Figure 1e. The first class is defined as 'cores'; these are sets of proteins that are present in most purifications of a complex, no matter which protein is tagged; they consisted on average of around three proteins, but ranged from one to 23 proteins. Altogether, Gavin *et al.* [1] found 491 complexes in yeast and an equivalent number of cores. In fact, they estimated that there may be up to 800 core machines in yeast. The second class comprises proteins often found together but not always with the same cores; such groups were called 'modules'. Gavin and colleagues identified 147 modules, of which 87 were mutually exclusive. Of the 87 modules, 31 appear to be related to differences in subcellular location, and might thus specify subtle differences in function. Most modules consisted of two or three proteins. Finally, a large number of proteins appear to be more or less loosely associated with cores and modules; these so-called 'attachments' may not always be essential for complex formation and may often represent modulators of the function of a protein complex. Interestingly, modules tend to be even more conserved, or

Table 2

Purifications leading to the definition of complex 314 in the study by Gavin *et al.* [1]

	Preys								
Baits	Aco1	Aha1	Aro1	Cct8	Gfa1	Pfk1	Pfk2	Yhb1	Additional
Aco1	-	-	-	-	-	-	-	-	-
Aha1	-	x	x	-	x	x	-	-	+26
Aro1	-	-	x	x	x	x	-	x	+24
Cct8	-	-	-	-	-	-	-	-	-
Gfa1	-	-	-	-	-	-	-	-	-
Pfk1	-	-	-	-	-	-	-	-	-
Pfk2	-	-	-	-	-	x	x	-	+3
Yhb1	x	-	-	-	x	-	-	x	+18
Prey count	10	7	17	10	106	30	21	13	

Purifications that lead to the definition of complex 314 as described by Gavin *et al.* [1] and in Figure 1e. Each line represents a single purification with the bait indicated. For each of the bait proteins, the columns indicate the prey proteins (x) associated with the bait or not found with the bait (-). Complex 314 consists of the core proteins Aro1 and Cct8, which were found in two purifications of which only one (bait: Aro1) is shown here. Aco1 was found in 10 purifications, that is, in 9 purifications in addition to the one using Yhb1 as bait. Only baits that were included in the 'final' complex by the SAI algorithm are shown in this table. In addition, the modules 103 (Pfk1-Pfk2, found in five purifications) and 114 (Aco1-Yhb1, found in three purifications) associated with the core, as indicated by the co-purification of Pfk2 with Pfk1 and of Aco1 with Yhb1. Finally, Aha1 and Gfa1 were classified as attachments because they were not found consistently associated with any of the other components and thus could not be classified as core or module. Note that Aha1, Aro1 and Yhb1 had many more proteins co-purified when they were used as baits than when they were prey. For example, when Aha1 was used as a bait, the four proteins Aha1, Aro1, Gfa1 and Pfk1 were identified as binding to it, plus another 26 proteins not shown here; these are indicated in the Additional column. Purifications with Aco1, Cct8, Gfa1 and Pfk1 as baits appear to have been unsuccessful. Note that the information in this table is not sufficient to derive the complex shown in Figure 1e but also requires information from additional purifications only indicated in the row 'Prey count' and in the column 'Additional'.

share the same function and localization, than are cores. Attachments often do not share a common function or localization although they appear to be well conserved.

These are not the first studies to get to grips with the yeast complexome. In the previous study from the EMBL and Cellzome authors [6], 1,739 proteins were tagged with TAP tags and the associated proteins analyzed. In another study, Ho *et al.* [7] tagged 725 proteins with the eight amino-acid FLAG epitope and purified the associated complexes. These datasets represent only subsets of the yeast proteome, however, and are only partially overlapping. For example, only 94 baits were common to both screens. Both groups also used quite different protocols for their analysis. Not surprisingly, the resulting complexes looked very different. On average, the number of proteins common to corresponding purifications was less than 9% of the total number of proteins in both datasets [8]. The degree of reproducibility was thus rather disappointing, even though it could be explained by the different protocols.

With the much more comparable procedures and comprehensive datasets from the two new studies, we can compare their results more rationally. Both groups tagged the vast majority of all yeast proteins, although only a third of these were ultimately purified, namely 1,993 in Gavin *et al.* [1] and 2,357 in Krogan *et al.* [2] (Table 1). While this does not sound a lot, most of these purifications co-purified with at least one interacting protein (namely 1,754 out of 1,993 attempts in [1]; no such number was given in [2]). Altogether, about 2,700 unique proteins were reliably identified this way by each group, corresponding to about 60% of the yeast proteome.

Gavin *et al.* [1] found 73% of the complexes that have been documented in the Munich Information Center for Protein Sequence (MIPS) database [9] (217 complexes) and the literature (62 complexes not in MIPS). Thus the study was comprehensive, but also missed many complexes. In fact, the authors mention that they have not found 74 complexes that have been reported in the literature. This may be due to technical limitations (for example, when membrane-associated complexes were involved) or to biological reasons (for example, because complexes form only under conditions not tested). On the other hand, 257 of the 491 complexes were entirely novel and only 20 of those known previously had no novel component in this study [1].

How do the two screens compare?

The two datasets cannot in fact be compared easily because of different data formats and computational methods used

to infer complexes from raw purification data. Gavin *et al.* [1] provide a list of baits and co-purifying preys, whereas Krogan *et al.* [2] do not show their raw purification data (instead, they provide four lists of interactions computationally generated from raw data). Both groups condense their raw purification data into one list of 'complexes' - in each case it is important to remember that these complexes do not necessarily correspond to real physical entities, but rather to perceived complexes (see Figures 1 and 2).

In the following discussion we will consider only these two lists of derived complexes. It is impossible to say which is of 'better quality' until the two raw datasets are systematically compared to thoroughly studied individual complexes (which will then serve as 'gold standards'). Also, both groups have applied various computational strategies to weed out false positives from their final complexes, which in turn affects the size of the complexes: the more stringent the weeding the fewer false positives there are, but the resulting complex may also have lost some biologically relevant proteins. That said, each group identified parameters that appear to represent a reasonable balance between removal of false positives and loss of real positives.

The 491 complexes found by Gavin *et al.* [1] comprise 1,483 proteins (including modules and attachments) or 23% of the yeast proteome, while the 547 complexes found by Krogan *et al.* [2] contain 2,702 proteins or 42% of the yeast proteome. When both datasets are combined they add up to 3,033 proteins or 47% of the yeast proteome. Interestingly, the intersection of both datasets contains only 1,152 proteins (18%). Given this overlap, it is a reasonable assumption that there are 800 to 900 complexes in yeast.

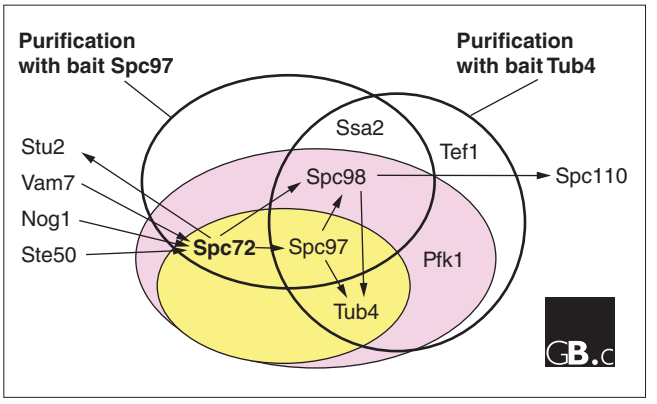

Figure 2
Protein complexes, purification data and two-hybrid interactions. Protein complex 285 from Krogan *et al.* [2] (composed of Spc72, Spc97 and Tub4 - the proteins in the area shaded yellow) was compared with complex 219 from Gavin *et al.* [1] (composed of all the proteins in the yellow and pink areas). Note that the two 'complexes' have been derived from many purifications by computational means and do not necessarily represent physical entities. The purifications from [1] using Spc97 and Tub4 as baits did not produce completely overlapping prey sets (there was no purification reported with Spc72 as bait); for example, only Tub4 but not Spc97 co-purified with Tef1 and Pfk1, whereas Spc97 but not Tub4 pulled down Spc72. Independently, two-hybrid screens have found a number of interactions between the members of these complexes and with other proteins, as indicated by arrows (pointing from the bait to the prey). The two-hybrid data are from [11,18-20].

Only six complexes are identical between the two datasets. Remarkably, 132 cores (27.62%) from the study of Gavin *et al.* [1] are completely contained in 115 complexes (21.02%) from the study by Krogan *et al.* [2], with an average overlap of 2.64 proteins. We found 188 complexes in [2] that do not share a single subunit with any complex found in [1]; by contrast, there are only 20 complexes in [1] which do not share any subunits with any of the complexes in [2]. A comparison of the two datasets is shown in Figure 3. Although our initial comparisons provide reasonable evidence that the two datasets are quite different, both groups need to run their own algorithm on the dataset of their competitor and see if they retrieve the same lists of complexes as with their own raw data. This would allow a comparison not only of the derived complexes but also of the underlying algorithms.

Comparison of protein purification and yeast two-hybrid data

It is difficult enough to compare the two datasets of complexes in [1] and [2], but it is even more difficult to compare them with protein-interaction datasets obtained with other methods. After complex purification, the most common procedure for identifying protein interactions is the yeast two-hybrid system [10], which discovers binary interactions but not complexes. Ideally, a two-hybrid screen using all the proteins of a complex would yield all the binary interactions within that complex, but this is rarely the case (Figure 2). In most cases, only a few interactions are discovered. On the other hand, the two-hybrid system often picks up weak interactions that are lost during complex purification because of the necessary washing steps. Thus, the data generated by protein-complex purification and two-hybrid analysis overlap even less than datasets obtained using the same method.

Comparison is also limited by the fact that no two-hybrid screen has been done in yeast that is as comprehensive as the protein-complex purification studies in [1] and [2]. Although the two-hybrid screens by Ito *et al.* [11] and by one of us (P.U.) and colleagues [12] claim to be comprehensive, they were by no means saturated. In fact, we estimate that only about 20% of the yeast genome has been used as baits and exhaustively screened by two-hybrid methods. In addition, two-hybrid screens suffer from a similar problem as protein-complex purifications: only about half of all screens yield reproducible interactions ([12] and C. Ester, R. Häuser, T. Kuhn, C. Müller, S.V. Rajagopala, B. Titz, P.U. and K. Wohlbold, unpublished observations). For example, we

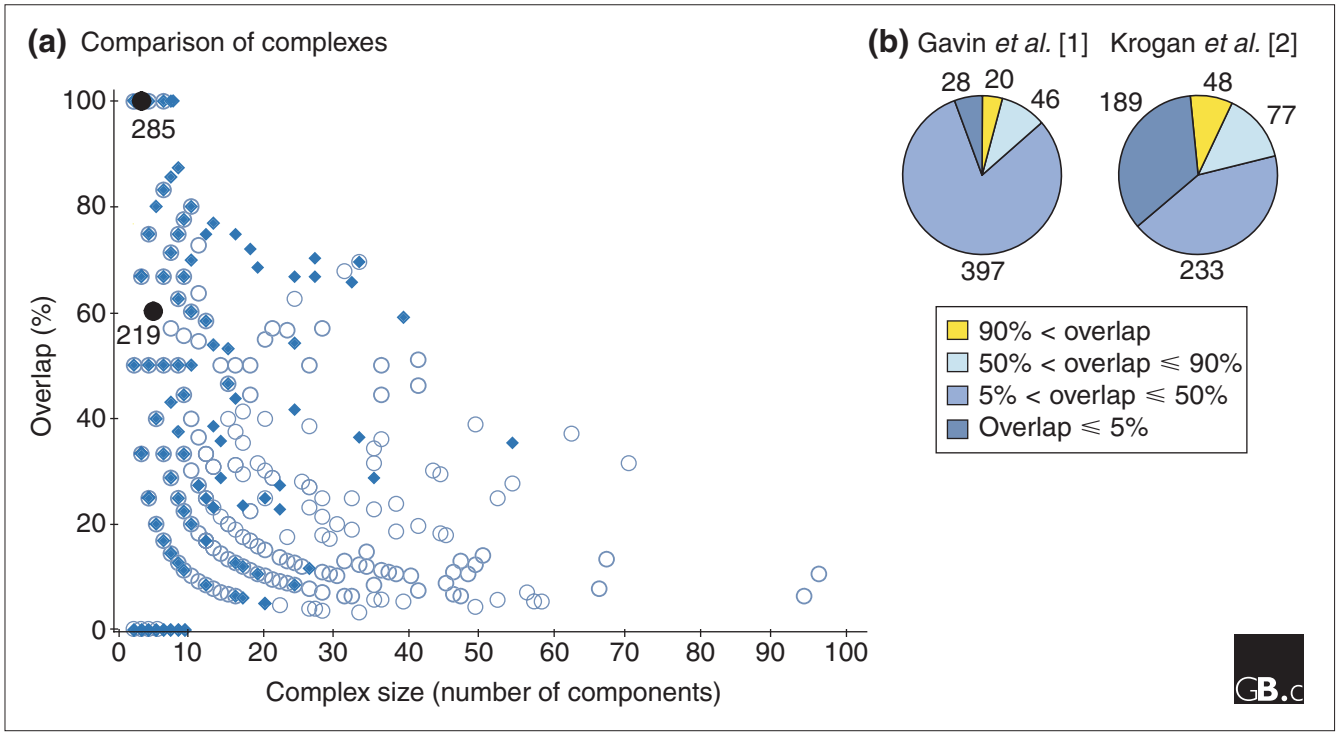

Figure 3
The two TAP studies yield dramatically different protein complexes. **(a)** Comparison of the composition of most similar complexes from Gavin *et al.* [1] and Krogan *et al.* [2]. Each circle represents a comparison of a complex from [1] to its most similar counterpart (that is, the complex that shares most of its proteins) in the dataset of [2]. The size of the complex is plotted along the *x*-axis and the percentage overlap in composition with the complex from [2] is plotted along the *y*-axis. The diamonds represent the same exercise carried out for complexes from [2] compared with their most similar counterparts from [1]. As an example, complex 219 (solid black circle) is 60% identical to complex 285 (solid black circle), whereas complex 285 has a 100% overlap with 219 (Figure 2). As each symbol refers to a one-way comparison, symbols may be superimposed but usually refer to different complexes/comparisons. **(b)** The pie-chart on the left shows the overlap between the 491 complexes identified in Gavin *et al.* [1], including cores, modules and attachments, and the 547 complexes from Krogan *et al.* [2], whereas the pie-chart on the right presents the converse analysis. For example, from the right-hand chart, there are 77 complexes reported in [2] that each have 50-90% of their proteins contained in a complex reported in [1]. In the left-hand chart, complex 219 from [1], with an overlap of 60% with complex 285 from [2] would be one of the 46 complexes in the slice showing the overlap range 50-90%. Conversely, complex 285 is one of the 48 complexes from [2] that overlap >90% with one of the complexes from [1], in this case complex 219. More details, such as which complexes are related to which, are available at [21].

found only 19 complexes in the dataset in [2] and 40 in that in [1] in which all proteins had previously been screened productively in two-hybrid screens. Most of these complexes are small, containing only two to five proteins. An example is shown in Figure 2.

Two-hybrid screens clearly do yield quite different interactions from protein-complex purifications. Given the very different nature of the methods this is hardly surprising. In fact, Aloy and Russell [13] have shown that protein purifications tend to pick up stable interactions whereas two-hybrid screens have a certain preference for transient interactions. It will be interesting to see how strong these trends are when truly quantitative and structural data become available. We have not compared the studies in [1] and [2] with other large-scale datasets such as genetic synthetic lethal screens, but such analyses will certainly be published shortly. For further comparisons with two-hybrid datasets or protein array data we will need more complete data. Comprehensive datasets using protein [14] or peptide arrays [15] are not available for yeast, but it is clear that they will also yield different results [16].

What remains to be done?

Gavin *et al.* [1] and Krogan *et al.* [2] have provided us with a glimpse of what the yeast complexome looks like in a mixture of happily growing cells. This is only half the truth. In nature, yeast is mostly starving and exposed to a variety of environmental conditions from heat to cold and wet to dry. We know that many physiological processes adapt with dramatic changes to such different growth conditions, and protein interactions reflect that. It would be exciting to see how the interactome reacts to such environmental factors, but such studies require much extra effort. Not only the interactome is subject to environmental influences: gene

expression, signal transduction and metabolism are all affected as well. Given that at least several thousand proteins appear to be phosphorylated and dephosphorylated in yeast [17], we begin to sense how complex even simple cells must be.

Comparative studies tell us that each analytic method only provides part of the truth. Although there are comprehensive datasets for purified complexes, there are only partial data for two-hybrid interactions and we have not even started to seriously apply protein arrays or structural genomics to the whole proteome or interactome of yeast. Let us not even think about more complex organisms.

Even assuming all those datasets had been collected under all conditions for all proteins and other compounds in a cell, and that we even knew how those molecules behave in space and time. Do we understand the cell? Not unless we can represent this plethora of information in computer-readable databases and information systems that can been understood by humans. Only if we manage to solve these informational problems as well as the technological ones will we be doing systems biology.

Acknowledgements

We thank Patrick Aloy, Rob Russell, and Anne-Claude Gavin for comments on an earlier version of the manuscript.

References

1. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, *et al.*: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440:**631-636.
2. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, *et al.*: **Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*.** *Nature* 2006, **440:**637-643.
3. **Protein complexes in yeast** [http://yeast-complexes.embl.de]
4 **The TAP project** [http://tap.med.utoronto.ca]
5. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422:**198-207.
6. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
7. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415:**180-183.
8. Cornell M, Paton NW, Oliver SG: **A critical and integrated view of the yeast interactome.** *Comp Funct Genom* 2004, **5:**382-402.
9. **Munich Information Center on Protein Sequences (MIPS)** [http://mips.gsf.de]
10. Fields S, Song O: **A novel genetic system to detect protein-protein interactions.** *Nature* 1989, **340:**245-246.
11. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
12. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403:**623-627.
13. Aloy P, Russell RB: **The third dimension for protein interactions and complexes.** *Trends Biochem Sci* 2002, **27:**633-638.
14. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, *et al.*: **Global analysis of protein activities using proteome chips.** *Science* 2001, **293:**2101-2105.
15. Jones RB, Gordus A, Krall JA, MacBeath G: **A quantitative protein interaction network for the ErbB receptors using protein microarrays.** *Nature* 2005, **439:**168-174.
16. Uetz P, Stagljar I: **The interactome of human EGF/ErbB receptors.** *Mol Systems Biol* 2006, **2:**E1-E2. doi 10.103814100048.
17. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R *et al.*: **Global analysis of protein phosphorylation in yeast.** *Nature* 2005, **438:**679-684.
18. Chen XP, Yin H, Huffaker TC: **The yeast spindle pole body component Spc72p interacts with Stu2p and is required for proper microtubule assembly.** *J Cell Biol* 1998, **141:**1169-1179.
19. Knop M, Pereira G, Geissler S, Grein K, Schiebel E: **The spindle pole body component Spc97p interacts with the gamma-tubulin of *Saccharomyces cerevisiae* and functions in microtubule organization and spindle pole body duplication.** *EMBO J* 1997, **16:**1550-1564.
20. Knop M, Schiebel E: **Spc98p and Spc97p of the yeast gamma-tubulin complex mediate binding to the spindle pole body via their interaction with Spc110p.** *EMBO J* 1997, **16:**6985-6995.
21. **Protein complexes in yeast - a comparative look at different datasets** [http://uetz.fzk.de/yeast-complexes]