FEBS Letters 480 (2000) 32–36 FEBS 23895

Minireview

Genome-wide protein interaction maps using two-hybrid systems

Pierre Legrain*, Luc Selig

Hybrigenics, 180 Avenue Daumesnil, Paris 75012, France

Received 19 May 2000

Edited by Gianni Cesareni

Abstract Automated sequence technology has rendered functional biology amenable to genomic scale analysis. Among genome-wide exploratory approaches, the two-hybrid system in yeast (Y2H) has outranked other techniques because it is the system of choice to detect protein–protein interactions. Deciphering the cascade of binding events in a whole cell helps define signal transduction and metabolic pathways or enzymatic complexes. The function of proteins is eventually attributed through whole cell protein interaction maps where totally unknown proteins are partnered with fully annotated proteins belonging to the same functional category. Since its first description in the late 1980's, several versions of the Y2H have been developed in order to overcome the major limitations of the system, namely false positives and false negatives. Optimized versions have been recently applied at multi-molecular and genomic scale. These genome-wide surveys can be methodologically divided into two types of approaches: one either tests combinations of predefined polypeptides (the so-called matrix approach) using various short-cuts to speed up the process, or one screens with a given polypeptide (bait) for potential partners (preys) present in complex libraries of genomic or complementary DNA (library screening). In the former strategy, one tests what one knows, for example pair-wise interactions between full-length open reading frames from recently sequenced and annotated genomes. Although based on a one-by-one scheme, this method is reported to be amenable to large-scale genomics thanks to multicloning strategies and to the use of small robotics workstations. In the latter, highly complex cDNA or genomic libraries of protein domains can be screened to saturation with high-throughput screening systems allowing the discovery of yet unidentified proteins. Both approaches have strengths and drawbacks that will be discussed here. None yields a full proteome-wide screening since certain proteins (e.g. some transcription factors) are not usable in Y2H. Novel two-hybrid assays have been recently described in bacteria. Applications of these time- and cost-effective assays to genomic screening will be discussed and compared to the Y2H technology.

Key words: Biological network; Pathway; Library screening; Protein array; Two-hybrid

1. The yeast two-hybrid system

The basic concept of the yeast two-hybrid system is to detect the interaction between two proteins via transcriptional activation of one or several reporter genes [1]. A classical eukaryotic transcription activator contains a domain that specifically binds to DNA sequences (the binding domain, BD) and a domain that recruits the transcription machinery (the activation domain, AD). In the two-hybrid system, these two domains are distinct polypeptides, each fused to a polypeptide, X and Y respectively (see Fig. 1). The basis of the assay is that transcription will occur only if X and Y interact together. This is an indirect genetic assay prone to false positive and false negative results that will be discussed below.

Many variations of this assay have been described, playing around with the choice of BD and AD sequences, copy number of the plasmids encoding these sequences, strength of promoters, nature of selectable markers and also the nature of reporter genes [2]. X and Y polypeptides are also from various sources, from prokaryotic or eukaryotic organisms and even from artificial sequences.

The yeast two-hybrid assay has been used for detecting interactions between two known proteins or polypeptides and also for searching for unknown partners (prey) of a given protein (bait). For technical reasons, baits are usually proteins fused to the BD polypeptide while prey are proteins fused to the AD polypeptide. In the former approach, false negative results are often encountered (i.e. two proteins known to interact together are not detected in the yeast two-hybrid assay), and after several negative attempts with various experimental settings, usually researchers moved to other systems to confirm the interaction or to analyze mutants, such as co-immunoprecipitation or column affinity chromatography. In the latter approach, most often a bait protein finds prey candidates – sometimes many – and those might have to be considered as potential false positive. In this case, the first question to address is the selectivity of the screen, i.e. the number of hits for the number of interactions tested. This number gives an idea on the capacity of the bait protein to activate transcription on its own. Similarly, prey proteins might be found regardless of specific interactions with the bait. These prey proteins are often referred to as 'sticky' proteins and should be discarded as dubious partners. In summary, a standard two-hybrid assay cannot take into account the specificity of all protein–protein interactions and will always give rise to a certain proportion of false positive and false negative results.

The availability of fully sequenced genomes, both of prokaryotic and eukaryotic organisms has led to large-scale studies on gene expression (functional genomics) and more recently on the proteome [3]. It has also been tempting to envisage large-scale studies for protein–protein interactions to complete exhaustive protein interaction maps ('interactome'). Variations of the yeast two-hybrid technology have been developed which are derived from the assays described

*Corresponding author.
E-mail: plegrain@hybrigenics.fr

PII: S0014-5793(00)01774-9

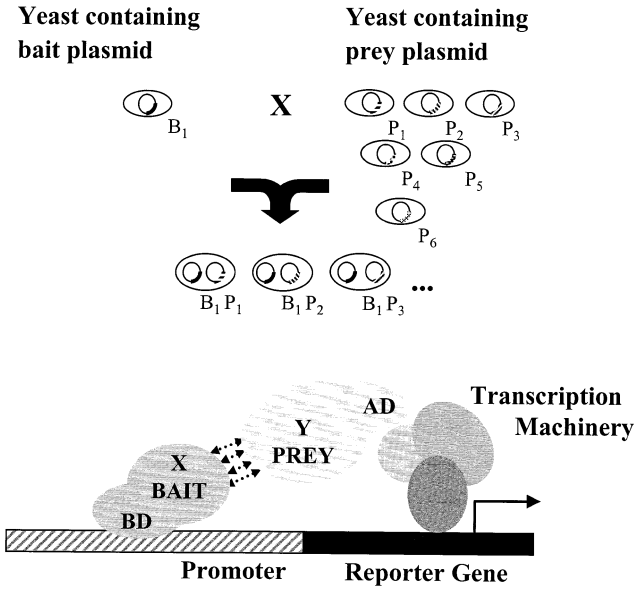


Fig. 1. General overview of the two-hybrid system by mating.

above. In many cases, protocols have taken advantage of the fact that yeast cells are haploid or diploid. When haploid yeast cells are of two mating types, they can fuse to form diploid cells. The two proteins of interest to be tested are produced separately in haploid cells of opposite mating type. When mixed together, diploid cells in which the two chimeric proteins are produced are formed, and activation of the diploid cell reporter genes is triggered upon interaction of the two proteins of interest [4].

2. The matrix approach

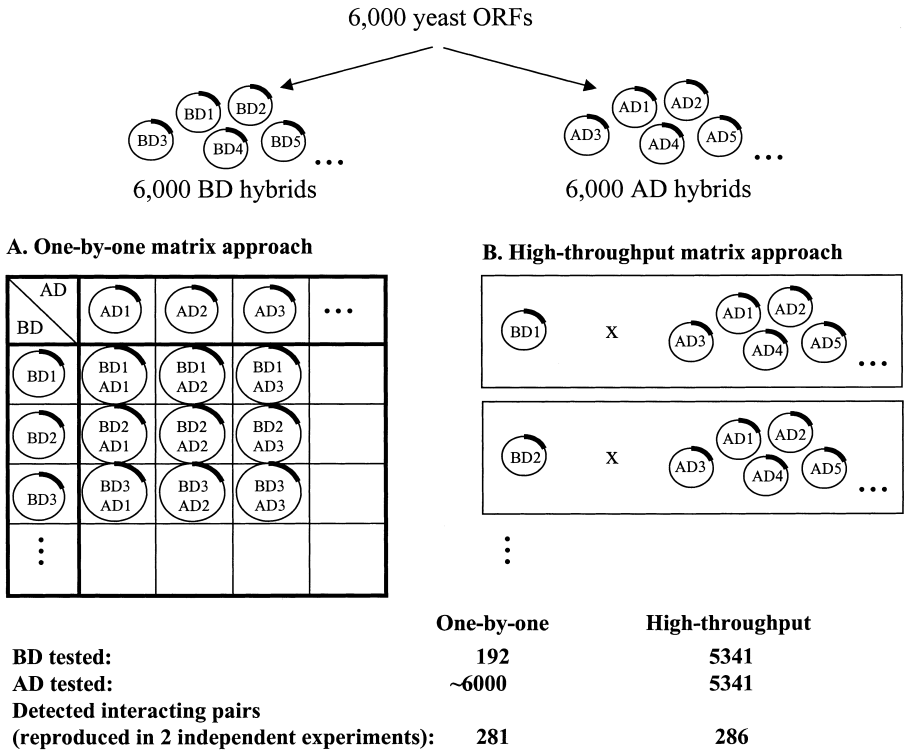
The matrix method was first described to explore interactions between *Drosophila* cell cycle regulators [5] and has been suggested to be applicable at a genome-wide scale to determine protein networks. Such an attempt was performed for the T7 phage proteome which contains 55 proteins [6].

As full genome sequences are publicly released, complete sets of coding sequences become available for cloning and testing in various assays, including the yeast two-hybrid. Two very recent papers have described large-scale approaches in the detection of many protein–protein interactions in the yeast proteome using sets of predefined open reading frames (ORFs) [7,8]. The ultimate aim would be to test all possible combinations between annotated ORFs of *Saccharomyces cerevisiae* (i.e. 4×10^7 combinations). ORFs are first amplified with specific primers, then a secondary PCR amplification is performed with common primers. In one case [7], these PCR products are cloned into BD and AD vectors and then transformed in yeast cells of opposite mating type. In the second strategy [8], PCR products are co-transformed with linear plasmids into yeast cells and gap repair occurs in vivo. Ultimately, yeast cells transformed with BD fusion plasmids or AD fusion plasmids are collected, stored and assayed individually or in pools. The common intrinsic limitation of this strategy is to test only full-length proteins that are predefined. However these different studies led to very different results that are summarized and discussed below.

In the publication by Uetz et al. [8], two strategies were followed, a time- and labor-intensive one and a high-throughput one, as defined by the authors themselves (see Fig. 2). In the first, a protein array of AD hybrids are produced separately in yeast cells grown in 384-well plates, these cells are mated individually with yeast cells transformed with a single BD plasmid and resulting diploid cells are selected for reporter gene activation. Experiments were performed with 192 different BD fusion proteins, giving rise to a range of 1–30 positive combinations (out of roughly 6000 assayed each time). Experiments were done twice and only interactors found in both experiments were considered as positives (20%). Reproducible interactions were found for only 87 BD fusion proteins that identified a total of 281 interacting protein pairs. This suggests a mean value of three interactions per protein. To achieve the second, a high-throughput approach, the authors made a pool of cells transformed with AD plasmids, mated them with cells transformed with one given BD plasmid and directly selected for the interactions. Out of a total of more than 5300 ORFs tested, only 817 were identified in putative protein–protein interactions, thus identifying a grand total of 692 interacting pairs, 59% of which were not reproducible (i.e. not found twice in two experiments). When this high-throughput approach is compared to the first one, only 12 out of the 87 BD proteins, previously selected as interacting partners, were also found engaged in interaction. This suggests that the high-throughput strategy increases considerably the number of false negatives that can be estimated to be above 90%, based on a mean value of three interactions per protein expected from the protein array approach (a total of 692 interactions for 5341 proteins tested). In addition, a high level of false positive interactions were probably retained since interacting pairs that were not found reproducibly were integrated into the results.

In the publication by Ito et al. [7], two collections of BD and AD plasmid transformed yeast cells were made and respectively pooled by groups of 96 clones. Before constituting the pools, any clone that activated the transcription of reporter genes on its own was discarded. In addition, in preliminary mating experiments, additional BD harboring clones that were clearly selected for interaction regardless of their partners were also discarded and new cleaned pools were constituted. Ultimately, 430 mating reactions were performed, leading to the analysis of more than 4 000 000 combinations (one tenth of a complete proteome analysis): 866 colonies were obtained, and 750 were successfully sequenced for DB and AD fusion plasmids. This extensive matrix identified 175 pairs of proteins (eight found bidirectionally), 12 of them being already known. This strategy clearly applies a strict selective pressure, avoiding many false positives, but this leads ultimately to very few interacting proteins (roughly 0.3 interactions per protein), close to what was detected in the previous high-throughput approach.

In conclusion, matrix approaches using predefined ORFs and a common assay for the detection of any pair-wise interaction in the proteome are prone to a very high level of false negative results. The rate of false positive interactions is more difficult to evaluate and is largely dependent on the criteria applied for the significance of the interactions i.e. the reproducibility of results in two experiments (see above) and the elimination of auto-activating clones. From the experiment with the protein array of AD proteins [8] and taking into account the results from the 87 BD proteins for which the



	One-by-one	High-throughput
BD tested:	**192**	**5341**
AD tested:	**~6000**	**5341**
Detected interacting pairs (reproduced in 2 independent experiments):	**281**	**286**

Fig. 2. Schematic overview of the two-hybrid matrix approaches. Two experimental settings are depicted. In the case of the HTS approach, 406 additional interactions were detected but not reproduced in a second independent experiment.

assay did work, one can extrapolate the total number of protein–protein interactions for the yeast proteome to be between 15 000 and 20 000 interactions. Any reliable proteome-wide strategy should aim for the detection of at least a reasonable fraction of these interactions, without retaining too many false positive results. Note, however, that a large number of the detected interactions in the present studies are new to the scientific community or involve proteins with unknown functions. Biological validation and/or integration of data from other sources will help predict the biological relevance of these interactions [9].

3. Library screening

Although primarily designed to detect a physical association between two known proteins, the Y2H assay became rapidly the most widely used system to screen libraries for the identification of interacting proteins (prey) with a known protein (bait). Further experiments were designed according to the availability of sequence and functional data on the prey protein to validate the interaction. However, as soon as the library was available, it was tempting to repeat such Y2H library screening experiments with proteins involved in the same biochemical process or with the prey protein used in turn as a bait protein. This led to the concept of specific functional protein interaction maps that could suggest a biological function for proteins never studied until then. One step further was reached when these experimental settings were designed as a proof of concept for proteome-wide approaches.

Several years ago, Fromont and colleagues generated a highly complex library of random yeast genomic fragments (over 3 000 000 fragments for a genome size of 15 000 000 base pairs) cloned into an AD vector [10]. This library was further transformed into yeast haploid cells, transformants were collected, pooled and frozen into multiple aliquots, each of these representative of the complete original library. These were used in subsequent screens to ensure reproducibility of each screening experiment. BD bait plasmids were prepared and transformed into yeast haploid cells of the opposite mating type. An efficient mating strategy was developed in order to achieve a full coverage of the library (over 30 000 000 diploids have to be produced in each screening experiment). Diploid cells producing BD and AD proteins were further plated on selective medium. This selective pressure was chosen according to the relative transcriptional activation potential of the BD protein. All selected positive colonies were identified by sequencing the AD plasmid. Prey proteins were classified according to their heuristic value based on statistical occurrence of the selected genomic fragment in the library. The most convincing prey proteins (with the highest heuristic values) could then serve as bait proteins in iterative screens. In this pilot study, a total of 15 screens was performed, 170 interactions were selected connecting 145 different yeast proteins. Among these interactions, 87 were of high and medium heuristic values leading to a mean value of 5.8 interactions per bait protein.

Another advantage of screening randomly generated fragments is the subsequent and immediate determination of interacting domains. This is well exemplified in a study on two interacting proteins involved in nuclear export for which a functional interacting domain was directly mapped along the screening experiment that identified the interacting protein [11] while another group had to postulate the interaction and to make time-consuming deletion experiments to map the very same domain [12]. The common sequence shared by the selected overlapping prey fragments defined experimen-

tally the smallest docking site of the prey, thus allowing a precise mapping of a functional domain.

This strategy has now been applied to many proteins in yeast. Based on this approach as well as genetics studies, a model of the RNA polymerase III preinitiation complex has been proposed [13]. Domains of interaction were defined for many components of the complex, filling gaps between 3D structures of monomers and the functional definition of the active complex. More than 100 yeast proteins involved in RNA metabolism have been screened for protein interactions leading to a network of interactions involving several hundreds of proteins [14].

This library screening approach has also been used for organisms other than yeast. Recently a large-scale study for protein–protein interactions involved in vulval development in *Caenorhabditis elegans* was published [15]. This study combines a matrix strategy for a set of 29 proteins known to be implicated in this developmental pathway and a library screening with these proteins to identify new players in the game. Another recent study deals with hepatitis C viral polypeptide interactions [16]. In this case the genome encodes a single polyprotein that is later processed by cellular and virally encoded proteases into 10 polypeptides. A matrix approach using the 10 canonical full-length mature polypeptides failed to detect any interaction between HCV polypeptides, including the well-known capsid oligomer or the heterodimer between the NS3 protease and its cofactor NS4A. This suggests again that pre-defined fusion proteins often present incorrect folding, expression, stability, or localization in the nucleus. On the contrary, exhaustive screening for interaction of randomly generated HCV genomic fragments revealed the expected capsid homodimer and viral protease heterodimer, as well as novel interactions. HCV polypeptides selected as prey will be advantageously used in further biochemical or genetics studies including the screen of human interacting proteins. Finally, an exhaustive proteome-wide approach for building the protein interaction map in *Helicobacter pylori* is in progress (J.C. Rain, PL et al., manuscript in preparation). This map will link half of the proteins of the proteome in a comprehensive network of protein–protein interactions.

Intrinsic limitations of the conventional Y2H system include its reliance upon transcriptional activation. The bait or the prey proteins may be capable of activating transcription by themselves. Besides, both BD and AD chimeric proteins need to be localized to the nucleus in order to trigger transcriptional activation. Other potential limitations include the absence in yeast of certain types of post-translational modifications that could be required to detect interactions between higher eukaryotic proteins. All of these major drawbacks of the yeast have been circumvented in novel Y2H systems [17–23], but still need to be adapted for the selection of novel interacting proteins in screening experiments.

4. Bacterial two-hybrids (B2H) as complementary approaches

Alternative approaches have been developed in other organisms than yeast. On the basis of throughput and cost of genome-wide studies, *Escherichia coli* appears as a more suitable host than *S. cerevisiae* because the generation time is much lower and molecular biology techniques are more adapted to bacteria than to yeast. Also, the high degree of competence of transformation of *E. coli* allows for the full coverage of highly complex genomic or cDNA libraries. Several B2H systems have been described to date, and can be split into two categories: those based upon transcriptional activation/repression of reporter genes or on reconstitution of an enzyme.

In the first category, the authors took advantage of the dimerization properties of the λ phage cI repressor [24–26] or the bacterial transcription repressor LexA [27]. cI and LexA repressor proteins can be divided into two functionally distinct domains, a N-terminal DNA BD unable to dimerize and a C-terminal domain which is strictly required for dimerization and therefore for the function of the repressors. This property permits the replacement of the C-terminal dimerization domain by virtually any heterologous homodimerizing domain. The first successful attempt with this system was to isolate mutations in the leucine zipper domain of the yeast transcription factor GCN4 that are critical for dimerization [24]. Lambda cI N-terminal fusions have also been used to study and quantify the homodimerization ability of various proteins [25]. In a first attempt to use this system as a screening tool, Bunker and Kingston [26] transformed a human cDNA library in bacteria already expressing cI-Myc chimeric proteins at low levels. The authors succeeded in isolating a cDNA whose product could compete with cI-Myc dimerization and therefore de-repress β-galactosidase production. This first attempt suffers from a major drawback, which is the intrinsic competition between homo- and hetero-chimeric dimers. In addition, no real selection method was developed to enable positive clones to grow. A novel LexA-based system demonstrated that heterodimerization could be examined in *E. coli* by the use of an hybrid operator allowing only heterodimers to bind [27]. Although suitable for the monitoring of heterodimer formation, this new system has however no obvious selection scheme which limits its use in library screening experiments. More universal *E. coli* two-hybrid systems, that could be used either for homo- or hetero-dimer detection, have been also described. In one experiment, Kornacker et al. [28] used the DNA binding properties of LexA and of the transcriptional activator AraC. The respective AraC and LexA operator sequences are placed flanking the reporter gene promoter. Upon binding of AraC-X and LexA-Y proteins to DNA, a DNA loop is induced (DNA bending) preventing expression of the reporter gene. 'Proof of concept' for screening of interacting partners was given when the Human Papilloma Virus (HPV) E6 protein encoding DNA was isolated from a complex DNA pool when the human E6AP protein – a known interactor of HPV E6 – was used as bait. As stated by the authors, introduction of a toxic gene in place of LacZ or the use of a toxic substrate for β-galactosidase is the obvious next step in order to select interacting partners from complex libraries. Another transcriptional activation assay, very similar to the Y2H system, has also been reported [29]. It shows that a protein–protein contact between a polypeptide bound to the DNA via a BD and another interacting polypeptide fused to RNA polymerase could trigger transcription. Again, experiments are underway to provide this system with a marker that would not only allow binders to be screened but also to be selected from a population of non-interacting proteins.

In the second category, Karimova et al. [30] described a novel type of bacterial two-hybrid system based on the reconstitution of a signal transduction pathway mediated by

cAMP. This system exploits the fact that the catalytic domain of the adenylate cyclase from *Bordetella pertussis* consists of two complementary fragments, namely T18 and T25, which are not active when physically separated. When these two fragments are fused to interacting polypeptides, X and Y, heterodimerization of these hybrid proteins results in functional complementation between T18 and T25 fragments and, therefore, cAMP synthesis. cAMP binds to the catabolite activator protein, CAP, which turns on the expression of several genes, including genes involved in lactose and maltose catabolism. In the presence of cAMP, bacteria therefore become able to utilize lactose or maltose as the unique carbon source and can be easily distinguished on indicator or selective media. The screening potential of this system on selective medium was evaluated in a *H. pylori* library screening experiment using a protein of *H. pylori* previously used in the Y2H system. Among the candidate colonies that were selected, two distinct families of preys were identified, one of which corresponds to a known interacting protein (L.S., P.L. et al., manuscript in preparation). Experiments are underway to derive new strains that could be used for high-throughput purposes. Pelletier et al. [31,32] described a B2H system based on the assembly and complementation of chimeric N- and C-terminal fragments of the murine dihydrofolate reductase (mDHFR). Endogenous *E. coli* DHFR can be inactivated by the antifolate drug trimethoprim, which has little effect on mammalian DHFR. DHFR being involved in biosynthesis of purines, thymidylate, methionine, and pantothenate, restoration of mDHFR activity upon interaction of the DHFR fusion proteins enabled cell propagation under trimethoprim low concentration. It was recently shown that this system could be used for selection of interactions within a collection of mutant proteins [32].

5. Conclusion

The recent development of these bacterial two-hybrid assays based on distinct experimental settings suggests that at least one of them should be available in the future for high-throughput screening (HTS) and proteome-wide analysis. Nevertheless, at the present time, the yeast two-hybrid assay remains the only large scale technology that is available to build protein interaction maps. Two strategies – namely the matrix approach and the library screening approach – are now being tested to find the most efficient way to explore proteomes for interactions (the interactome). It should be stressed that the library screening approach is much more selective than the matrix approach (in a classical library screening experiment several tens of millions combinations are assayed for an output of less than 1000 positive clones). Nevertheless, more interactions per bait protein are identified through library screening, reducing considerably the false negative rate. This is explained by the nature of the fragments that are tested. While only one fusion event is tested when full-length proteins are used in a matrix approach (eventually two when the reciprocal combination is assayed), several tens of overlapping domains of the same protein are tested for interaction in a highly complex library screening, thus increasing the chance that a given fragment will be active in the two-hybrid assay.

Reducing the false negative rate and evaluating the heuristic value of each detected interaction in order to decrease the rate of false positive results should lead to the building of accurate protein interaction maps that will be the basis for the exploration of novel protein function and pathways. The challenge will be then to provide a bioinformatics tool that allows the exploration of this protein interaction map, connected to any other functional annotation on the relevant proteins.

References

[1] Fields, S. and Song, O. (1989) Nature 340, 245–246.
[2] Vidal, M. and Legrain, P. (1999) Nucleic Acids Res. 27, 919–929.
[3] Goffeau, A. (2000) FEBS Lett., this issue.
[4] Bendixen, C., Gangloff, S. and Rothstein, R. (1994) Nucleic Acids Res. 22, 1778–1779.
[5] Finley Jr., R.L. and Brent, R. (1994) Proc. Natl. Acad. Sci. USA 91, 12980–12984.
[6] Bartel, P.L., Roecklein, J.A., SenGupta, D. and Fields, S. (1996) Nat. Genet. 12, 72–77.
[7] Ito, T. et al. (2000) Proc. Natl. Acad. Sci. USA 97, 1143–1147.
[8] Uetz, P. et al. (2000) Nature 403, 623–627.
[9] Tsoka, S. and Ouzounis, C.A. (2000) FEBS Lett. 480, 37–41.
[10] Fromont-Racine, M., Rain, J.C. and Legrain, P. (1997) Nat. Genet. 16, 277–282.
[11] Siomi, M.C., Fromont, M., Rain, J.C., Wan, L., Wang, F., Legrain, P. and Dreyfuss, G. (1998) Mol. Cell. Biol. 18, 4141–4148.
[12] Truant, R., Fridell, R.A., Benson, R.E., Bogerd, H. and Cullen, B.R. (1998) Mol. Cell. Biol. 18, 1449–1458.
[13] Flores, A. et al. (1999) Proc. Natl. Acad. Sci. USA 96, 7815–7820.
[14] Fromont-Racine, M. et al. (2000) in press.
[15] Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. and Vidal, M. (2000) Science 287, 116–122.
[16] Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspé, G., Tiollais, P., Transy, C. and Legrain, P. (2000) Gene 242, 369–379.
[17] Du, W., Vidal, M., Xie, J.E. and Dyson, N. (1996) Genes Dev. 10, 1206–1218.
[18] Aronheim, A., Zandi, E., Hennemann, H., Elledge, S.J. and Karin, M. (1997) Mol. Cell. Biol. 17, 3094–3102.
[19] Aronheim, A. (1997) Nucleic Acids Res. 25, 3373–3374.
[20] Marsolier, M.C., Prioleau, M.N. and Sentenac, A. (1997) J. Mol. Biol. 268, 243–249.
[21] Johnsson, N. and Varshavsky, A. (1994) Proc. Natl. Acad. Sci. USA 91, 10340–10344.
[22] Osborne, M.A., Dalton, S. and Kochan, J.P. (1995) Biotechnology (NY) 13, 1474–1478.
[23] Tirode, F., Malaguti, C., Romero, F., Attar, R., Camonis, J. and Egly, J.M. (1997) J. Biol. Chem. 272, 22995–22999.
[24] Hu, J.C., O'Shea, E.K., Kim, P.S. and Sauer, R.T. (1990) Science 250, 1400–1403.
[25] Di Lallo, G., Ghelardini, P. and Paolozzi, L. (1999) Microbiology 145, 1485–1490.
[26] Bunker, C.A. and Kingston, R.E. (1995) Nucleic Acids Res. 23, 269–276.
[27] Dmitrova, M., Younes-Cauet, G., Oertel-Buchheit, P., Porte, D., Schnarr, M. and Granger-Schnarr, M. (1998) Mol. Gen. Genet. 257, 205–212.
[28] Kornacker, M.G., Remsburg, B. and Menzel, R. (1998) Mol. Microbiol. 30, 615–624.
[29] Dove, S.L., Joung, J.K. and Hochschild, A. (1997) Nature 386, 627–630.
[30] Karimova, G., Pidoux, J., Ullmann, A. and Ladant, D. (1998) Proc. Natl. Acad. Sci. USA 95, 5752–5756.
[31] Pelletier, J.N., Campbell-Valois, F.X. and Michnick, S.W. (1998) Proc. Natl. Acad. Sci. USA 95, 12141–12146.
[32] Pelletier, J.N., Arndt, K.M., Pluckthun, A. and Michnick, S.W. (1999) Nat. Biotechnol. 17, 683–690.