*2464–2469* *Nucleic Acids Research, 1997, Vol. 25, No. 12* 

# Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins

**Siegfried Böhm*, Dmitrij Frishman[1] and H. Werner Mewes[1]**

Max-Delbrück-Centrum für Molekulare Medizin (MDC), Robert-Rössle-Straße 10, 13125 Berlin-Buch, Germany and [1]Max-Planck-Institut für Biochemie, MIPS, Am Klopferspitz 18a, 82152 Martinsried, Germany

Received January 29, 1997; Revised and Accepted April 21, 1997

## ABSTRACT

**The PROSITE pattern Zinc_Finger_C2H2 was extended to permit the detection of all C2H2 zinc fingers and their parent proteins in the recently completed sequence of the yeast genome. Additionally, a new computer program was written that extracts other zinc binding motifs (non C2H2 'fingers'), overlapping with the classical zinc finger pattern, from the found set of yeast C2H2 fingers. The complete and correct detection of all fingers is a prerequisite for the classification of the yeast zinc finger proteins in functional terms. The detected 53 yeast C2H2 zinc finger proteins do not contain finger clusters with 10 or more repeats, as is frequently found in higher eukaryotes. Only three proteins contain four or more fingers in a cluster. Moreover, nearly all 27 yeast proteins with tandem arrays of two or three finger domains can be classified into nine subgroups with high sequence conservation in their finger clusters, in particular of their DNA recognition helices. These results and application of the recently elaborated finger/DNA recognition rules suggest that the yeast proteins belonging to the same subgroup may recognize identical or very similar DNA sites.**

## INTRODUCTION

In 1985 three groups independently observed in the DNA/RNA binding transcription factor TFIIIA from *Xenopus laevis* (1) a 9-fold repeated pattern of amino acids with conserved cysteine, histidine and hydrophobic residues (2–4). The arrangement of this pattern in TFIIIA is \$-X-C-$X_{2,4,5}$-C-$X_3$-\$-$X_5$-\$-$X_2$-H-$X_{3,4}$-H, where X represents any amino acid, \$ a hydrophobic residue, C cysteine and H histidine. Based on this observation, as well as on biochemical and biophysical studies, Klug and co-workers (2) coined the term 'zinc finger' to describe their proposal that this ~30 amino acid sequence motif forms an independent DNA binding minidomain folded around a central zinc ion with tetrahedral arrangement of cysteine and histidine metal ligands (reviewed in 5). Through the tandem repetition of structurally identical small finger domains with chemically different DNA recognition parts (mainly the N-terminal half of an α-helix) truly modular recognition of specific DNA sites is facilitated in the zinc finger proteins. Crystal structures of zinc finger–DNA complexes, site-directed mutagenesis and screening/selection studies have revealed finger/DNA recognition rules (a code) that can describe, at least partially, the sequence-specific interactions between fingers and DNA and that are useful for the *de novo* design of zinc finger proteins that recognize desired DNA target sites (see 5–8 and references therein).

Zinc finger proteins represent perhaps the largest and most diverse superfamily of nucleic acid binding proteins in eukaryotes. It has been estimated that up to 1% of genes in the human genome may encode proteins with zinc finger domains (9). Our database of zinc finger proteins (Zfp) and zinc fingers (Zf) now contains >560/2800 Zfp/Zf entries. Since the earlier collection of Zfp/Zf was published by Jacobs (10) in 1992, the number of Zfp/Zf sequences has increased nearly 3-fold.

Sequence analysis of the fingers in our database has indicated that the PROSITE pattern Zinc_Finger_C2H2 (11) used for the detection of zinc fingers in new protein or genome sequences does not match all fingers actually present. The complete and correct detection of protein sequence motifs is essential in analysing the huge amount of data provided by the many genome sequencing projects. Very recently the genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced through a worldwide collaboration (12). The sequence of ~12 000 kb represents the first complete genome sequence for a eukaryote, defining 6305 potential protein encoding genes. Here we report on extensions of the PROSITE C2H2 pattern to permit the detection of all fingers and their parent proteins in the yeast genome. Moreover, applying a new program we have extracted from the set of detected yeast C2H2 finger sequences wrong 'fingers' which belong to other subfamilies of zinc binding motifs (13) but overlap partially with the classical finger motif. The implications of the detection of all C2H2 fingers for the functional classification of yeast Zfp are discussed for proteins with a finger pair in terms of their sequence-specific DNA recognition.

## MATERIALS AND METHODS

Using the SwissProt, PIR, EMBL and GenBank databases, literature searches and contributions from authors, we have created a Zfp/Zf

*To whom correspondence should be addressed. Tel: +49 30 9406 2478; Fax: +49 30 9406 2548; Email: boehm@mdc-berlin.de

database with >560/2800 entries. About half of the Zfp are complete sequences. The database includes an earlier collection of Jacobs (10), kindly supplied to us by the author. A total of 2475 non-identical finger sequences present in the whole Zf collection have been aligned using Clustal W (14) in 29 positions according to our new C2H2 pattern (see below and Fig. 1). The position-dependent frequencies of all 20 amino acids as well as a profile (15) were calculated from this non-redundant set of aligned finger sequences. Similar analyses were made with the complete set of 127 yeast fingers found with our new C2H2 pattern. In addition, a phylogenetic tree (16) was created from the distance matrix of the yeast fingers and used together with profile searches to qualify the sequences on a scale from 'good' to 'bad' fingers. A new program was written in the Perl programming language which permits the automatic detection of C2H2 fingers with our extended C2H2 pattern as well as the detection of 13 other known or putative zinc binding patterns (often described as 'finger' motifs) in protein sequences.

Searches with the program have been performed against the 6305 open reading frames (ORFs) of the translated complete yeast genome (12). (For sequence informations contact our home page, http://www.mips.biochem.mpg.de, or other yeast-related Internet resources, shown in table 1 of ref. 12.) The program for the automatic detection of zinc fingers in protein sequences can be obtained from one of us (D.F., frishman@mips.biochem.mpg.de). For information on the Zf/Zfp database and additional data not shown here contact the corresponding author (S.B., boehm@mdc-berlin.de).

## RESULTS AND DISCUSSION

### The C2H2 zinc finger pattern

Release 32 of the PROSITE (11) pattern ZINC_FINGER_C2H2 (accession no. PS00028), formulated as C-$X_{2,4}$-C-$X_3$-(F,Y,W,C,L,I, V,M)-$X_8$-H-$X_{3,4,5}$-H, contains 311 zinc finger proteins and 1350 zinc fingers, with 90 and 97% true positives respectively. A search against all yeast ORFs with the original PROSITE C2H2 zinc finger motif (further referred to as C2H2ori) resulted in 48/85 Zfp/Zf hits. A more detailed analysis of these yeast Zfp/Zf indicates that the above set is incomplete. In this work we extend the C2H2 pattern to permit the detection of all (except very unusual) yeast zinc fingers and their parent proteins. We introduced several modifications in the C2H2ori pattern and created two new extended patterns based on statistical analysis of 2475 non-identical finger sequences in our database (see Materials and Methods).

One of the varied patterns permits any amino acid in finger position 13 (see Fig. 1), which is restricted in C2H2ori to eight hydrophobic residues. In addition, two variable residues before the first invariant Cys have been allowed to match the whole fold of the finger domain (5). According to our analysis the mentioned eight hydrophobic residues occur in position 13 with a frequency of 95.6%, but all other residues except proline are also found in position 13. As a rule, the first position of aligned finger sequences, situated at the N-terminus of the first β-strand (Fig. 1), is also occupied by conserved hydrophobic residues (including in the calculation the frequently occurring His) with a frequency >95%. All other amino acid residue types occur rarely in position 1. In accord with these data, the first extended pattern, dubbed the canonical C2H2 motif (C2H2can), is therefore formulated as $X_2$-C-$X_{2,4}$-C-$X_{12}$-H-$X_{3,4,5}$-H (Fig. 1). Searches with the C2H2can

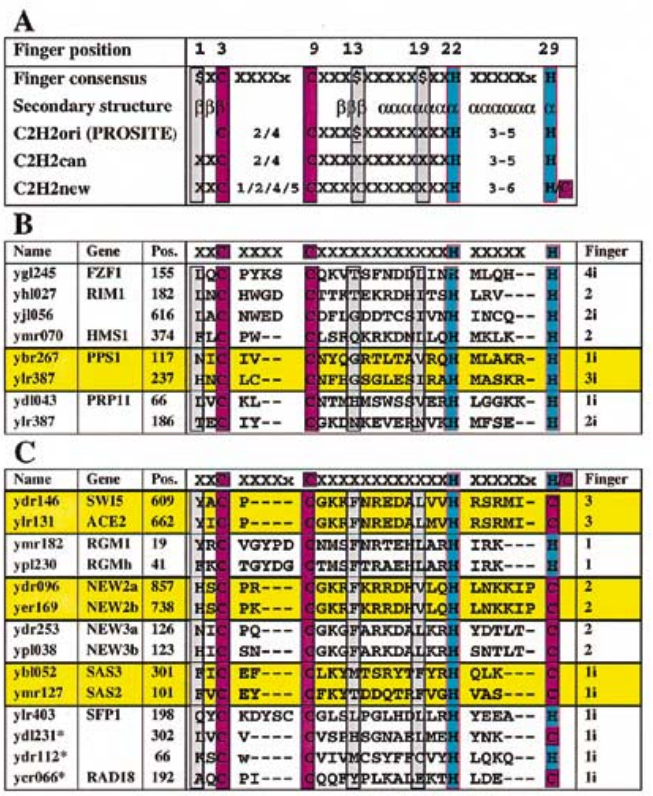



**Figure 1.** New C2H2 zinc fingers detected in the yeast genome. (**A**) Finger consensus, secondary structure and C2H2 motifs used in this work. The original PROSITE motif C2H2ori is shown together with the two extended patterns, C2H2can and C2H2new, proposed in this work. Position numbers of invariant zinc ligands and conserved hydrophobic residues in the zinc finger consensus according to the C2H2new pattern are shown in the first line. The finger consensus reflects 2475 non-identical sequences between positions 1 and 29 with invariant Cys and His in positions 3, 9 and 22 respectively and semi-invariant His/Cys in position 29. The conservation of hydrophobic residues in positions 1, 13 and 19 is ~95% (see text). Residue codes: C, cysteine; H, histidine; X, any amino acid; x, any amino acid in positions with rare occurance; $, hydrophobic residue, with one of FYWCLIVM in position 13 of the C2H2ori pattern. Secondary structure codes: α, α-helix; β, β-sheet. Numbers in the pattern show allowed spacings between the two pairs of invariant zinc ligands, where the / separator means one of and the – separator means the range. Colour code: purple, C; blue, H; grey, positions 1, 13 and 19. (**B**) Zinc fingers detected by the C2H2can pattern in addition to the C2H2ori motif. (**C**) Zinc fingers detected by the C2H2new pattern in addition to the C2H2can motif. In (B) and (C) the first three columns contain systematic yeast ORF names, gene names (where known) and location of the first amino acid of the zinc finger in the parent protein. Then follows the sequence of the zinc finger. The column marked Finger contains the numbering of the finger in the corresponding Zfp. Isolated fingers that are not part of a finger cluster are indicated by i (see legend to Fig. 3). Pairs of fingers with significant sequence similarities are boxed and/or coloured yellow. A name marked by * indicates a possible finger which might be a false positive detection.

pattern in the yeast genome led to detection of eight additional fingers compared with searches with C2H2ori (Figs 1 and 3). The second pattern further generalizes C2H2can to allow for the alternative presence of Cys in position 29 instead of the invariant His (Fig. 1), because our database analysis indicates ~3% of such altered finger types. Additionally, more flexible spacing has been permitted between the first invariant pair of zinc ligands (cysteines) and the second pair of zinc ligands (histidines or histidines/cysteines). The resulting C2H2new pattern can be considered as the least stringent C2H2 motif, formulated as $X_2$-C-$X_{1,2,4,5}$-C- $X_{12}$-H-

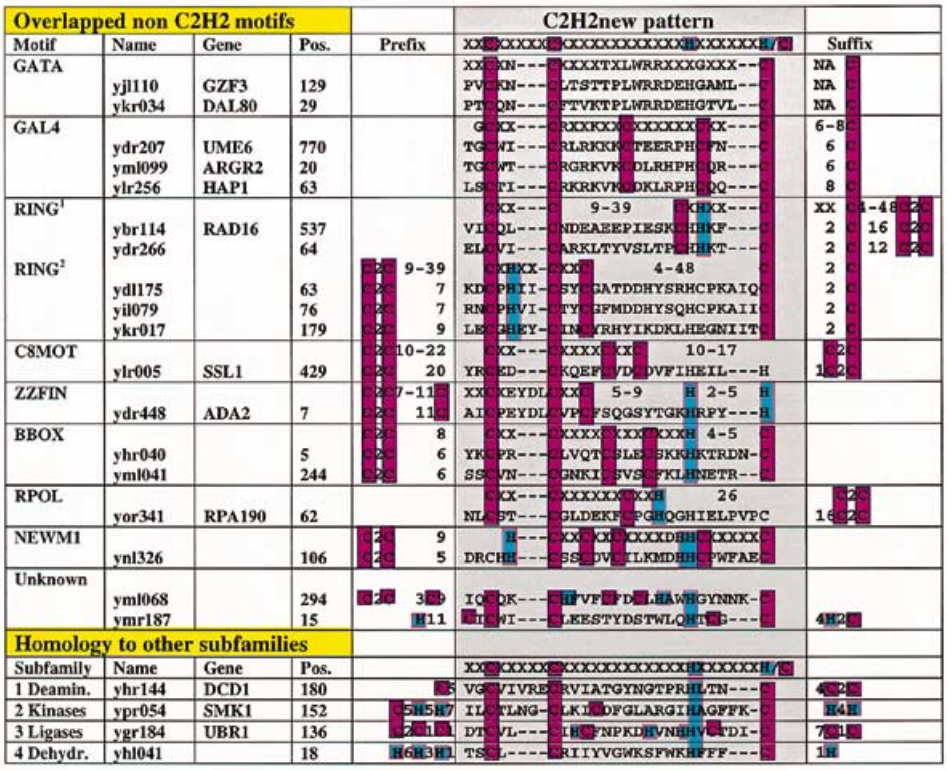


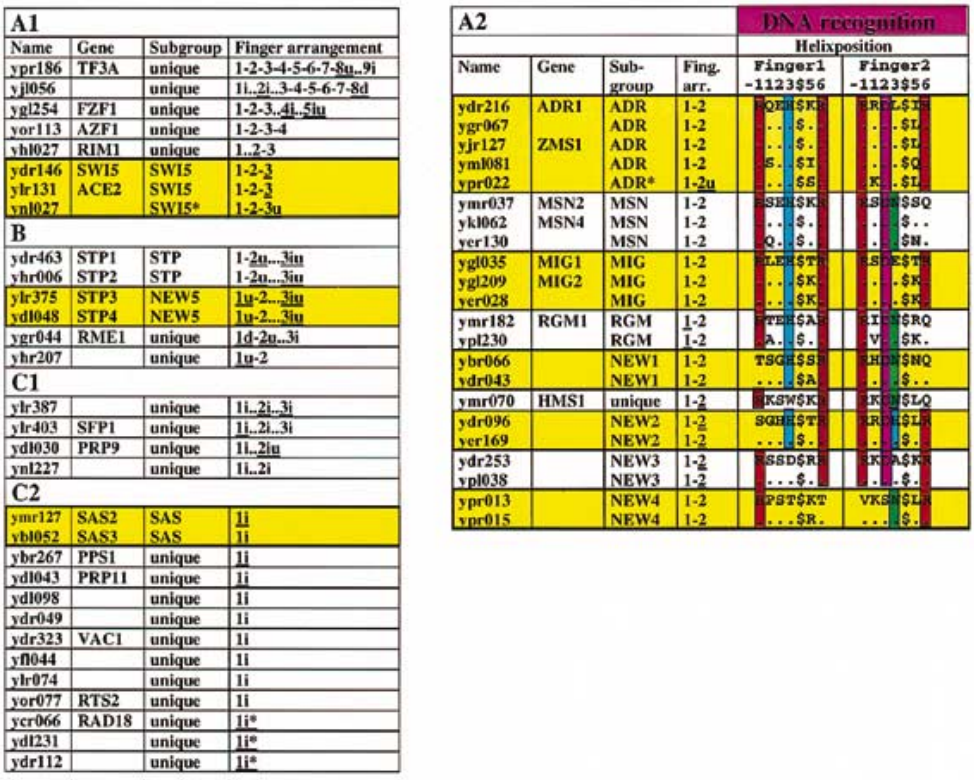

**Figure 2.** Wrong fingers overlapping with other non-C2H2 'finger' motifs and questionable fingers found in yeast proteins which belong to homologous non-Zfp subfamilies of proteins. Overlapping non-C2H2 motifs. In the first column the name of the non-C2H2 motif is shown; the next three columns contain the yeast ORF names, gene names and location of the 'finger' in the parent protein; the last three columns show the alignments of the overlapping C2H2new pattern sequence part with the non-C2H2 finger motifs (shown in each case in the first lines of the different motifs). Note that all non-C2H2 motifs have additional sequence parts to the left and/or right termini of their C2H2new part, marked with Prefix and Suffix. All numbers given in the sequence columns indicate spacings by amino acids of any type varying in length by the given numbers, X means any amino acid. RING[1] and RING[2] indicate two types of overlap. The colouring of C and H is as in Figure 1. Homology to subfamilies. In the first column a short name for the protein subfamily is given. All other columns are as described above. The short names refer to the following subfamilies, first described for the yeast proteins in the given references: 1 Deamin., cytidine deaminase (26); 2 Kinases, MAP kinase (27); 3 Ligases, UBR1 ligase (28); 4 Dehydr., NADH dehydrogenase.

$X_{3–6}$-(H,C). Applying the C2H2new motif to yeast genome searches increased the number of detected finger sequences by 34 compared with the results with the C2H2can motif (Figs 1–3). In summary, there is an increase by about one third (127 versus 85) of fingers detected with C2H2new compared with C2H2ori. Note that even with the C2H2new motif a few fingers with very unusual patterns are not found in the yeast genome, as well as in the whole Zf/Zfp database.

## Overlapping of different zinc 'finger' motifs

Searches with the C2H2ori and C2H2new motifs may lead in some cases to detection of questionable or even wrong fingers with sequences significantly deviating from a finger consensus. Compared with the C2H2ori motif the C2H2new pattern detects many more 'bad' fingers. Remarkably, nearly all (20 out of 22) 'bad' fingers (see Fig. 2) contain an alternative Cys instead of His in alignment position 29. Only a few such C2HC fingers have the profiles of 'good' fingers (Fig. 1). 'Bad' fingers may, for example, violate high conservation of hydrophobic residues in positions 1, 13 and 19 (Fig. 1), which is important for correct packing of the hydrophobic core of the finger domain (5). As described above for positions 1 and 13, a high conservation of hydrophobic residues (~96%) is also found in position 19. Moreover, in 'bad' fingers often other Cys and/or His residues are found in addition to the invariant Cys/His. These additional potential zinc ligands may occur in internal finger positions as well as in positions adjacent to their N- and/or C-termini (Fig. 2).

Our profile searches (15) and tree analysis (16) of the found 127 yeast fingers (data not shown) reveals that most of the 'bad' fingers are clearly separated from 'good' fingers through their low scores in profile searches and their large phylogenetic distances in the tree. In contrast, all 'good' fingers, almost exclusively present in Zfp with tandem arrays of two or more fingers (72 out of the 127 fingers), have the highest scores in the profile searches and are found in well-defined clusters with the lowest phylogenetic distances. Remarkably, 'bad' fingers are only found in Zfp with a single finger. Moreover, a separate sequence analysis of these 'bad' fingers has also indicated that in most cases their C2H2 pattern overlaps with motifs of other families of zinc binding domains (see Fig. 2), e.g. the RING-, GAL4- and GATA-type families of zinc binding domains (reviewed in 13). The overlap of several non-C2H2 finger motifs with the pattern of classical fingers may result in incorrect assignment of zinc binding motifs as C2H2 fingers. To overcome this complication we developed a computer program (see Materials and Methods) which compares a given C2H2 finger sequence with all available non-C2H2 patterns and detects overlaps. Among the 13 'non C2H2' motifs implemented until now in the program are the GAL4, GATA, LIM, steroid DBD, RING and TF2S zinc ribbon motifs taken



**A1**

| Name | Gene | Subgroup | Finger arrangement |
|---|---|---|---|
| ypr186 | TF3A | unique | 1-2-3-4-5-6-7-8u..9i |
| yjl056 | | unique | 1i..2i..3-4-5-6-7-8d |
| ygl254 | FZF1 | unique | 1-2-3..4i..5iu |
| yor113 | AZF1 | unique | 1-2-3-4 |
| yhl027 | RIM1 | unique | 1..2-3 |
| ydr146 | SWI5 | SWI5 | 1-2-3 |
| ylr131 | ACE2 | SWI5 | 1-2-3 |
| ynl027 | | SWI5* | 1-2-3u |
| **B** | | | |
| ydr463 | STP1 | STP | 1-2u...3iu |
| yhr006 | STP2 | STP | 1-2u...3iu |
| ylr375 | STP3 | NEW5 | 1u-2...3iu |
| ydl048 | STP4 | NEW5 | 1u-2...3iu |
| ygr044 | RME1 | unique | 1d-2u..3i |
| yhr207 | | unique | 1u-2 |
| **C1** | | | |
| ylr387 | | unique | 1i..2i..3i |
| ylr403 | SFP1 | unique | 1i..2i..3i |
| ydl030 | PRP9 | unique | 1i..2iu |
| ynl227 | | unique | 1i..2i |
| **C2** | | | |
| ymr127 | SAS2 | SAS | 1i |
| ybl052 | SAS3 | SAS | 1i |
| ybr267 | PPS1 | unique | 1i |
| ydl043 | PRP11 | unique | 1i |
| ydl098 | | unique | 1i |
| ydr049 | | unique | 1i |
| ydr323 | VAC1 | unique | 1i |
| yfl044 | | unique | 1i |
| ylr074 | | unique | 1i |
| yor077 | RTS2 | unique | 1i |
| ycr066 | RAD18 | unique | 1i* |
| ydl231 | | unique | 1i* |
| ydr112 | | unique | 1i* |

**A2**

| | | | | DNA recognition | |
|---|---|---|---|---|---|
| | | | | Helixposition | |
| Name | Gene | Sub-group | Fing. arr. | Finger1 -1123$56 | Finger2 -1123$56 |
| ydr216 | ADR1 | ADR | 1-2 | QEH$KR | RRL$TR |
| ygr067 | | ADR | 1-2 | ..$. | .$L |
| yjr127 | ZMS1 | ADR | 1-2 | ..$. | .$L |
| yml081 | | ADR | 1-2 | S.$I | .$Q |
| ypr022 | | ADR* | 1-2u | ..$S | K.$L |
| ymr037 | MSN2 | MSN | 1-2 | RSEH$KR | RSRH$SQ |
| ykl062 | MSN4 | MSN | 1-2 | ..$. | .$.. |
| yer130 | | MSN | 1-2 | Q.$. | .$N. |
| ygl035 | MIG1 | MIG | 1-2 | RLEK$TR | RSRE$TR |
| ygl209 | MIG2 | MIG | 1-2 | ..$K | .$K |
| yer028 | | MIG | 1-2 | ..$K | .$K |
| ymr182 | RGM1 | RGM | 1-2 | RTEH$AR | RIRH$RQ |
| ypl230 | | RGM | 1-2 | A.$. | V.$K. |
| ybr066 | | NEW1 | 1-2 | TSGH$SR | RHDH$NQ |
| ydr043 | | NEW1 | 1-2 | ...$A | .$.. |
| ymr070 | HMS1 | unique | 1-2 | RKSW$KR | RKRH$LQ |
| ydr096 | | NEW2 | 1-2 | SGHE$TR | RRRH$LR |
| yer169 | | NEW2 | 1-2 | ...$. | .$. |
| ydr253 | | NEW3 | 1-2 | RSSD$RR | RKRA$KR |
| ypl038 | | NEW3 | 1-2 | ...$. | .$. |
| ypr013 | | NEW4 | 1-2 | RPST$KT | VKSH$LR |
| ypr015 | | NEW4 | 1-2 | ...$R. | ...$. |

**Figure 3.** Complete set of classified zinc finger proteins in the yeast genome. The zinc finger proteins (Zfp) are classified into three subsets (A–C) according to the number, kind and arrangements of their fingers: (**A**) Zfp with tandem arrays of fingers in a cluster, with A1 containing proteins with three or more fingers and A2 containing proteins with a finger pair; (**B**) Zfp with one canonical and two unusual fingers in unusual arrangements; (**C**) Zfp with dispersed fingers (C1) or with a single finger (C2). The first two columns contain systematic yeast ORF names and gene names (where known). The next column describes the yeast subgroups defined by Zfp with identical numbers, patterns and arrangements of their fingers as well as homologous finger sequences. A subgroup name marked by * indicates a less closely related member in the considered subgroup. Subgroup names are coined as a rule from the experimentally best-characterized member of a subgroup. The subgroups NEW1–5 contain exclusively new ORFs not yet experimentally investigated. Unique Zfp have no homologous proteins in the yeast genome. Note that a few yeast Zfp (e.g. TF3A) which belong to subfamilies of Zfp conserved in evolution from lower to higher eukaryotes are not marked in this figure. In the column describing finger arrangement, additional fingers detected with the C2H2can or C2H2new motifs (see Fig. 1) are underlined, fingers with unusual patterns found only by visual inspection are indicated by u and underlined, d means a degenerate finger with one mutated zinc ligand and i stands for isolated (dispersed) fingers. 1* indicates a questionable finger. The linker sequence in finger tandem repeats with a consensus length of five residues (in a few cases with two, three or six residues) is given by (–), fingers linked by 10 or more residues are considered as dispersed or single fingers and are connected by (.. or ...). In the last two columns the sequences of the finger DNA recognition helices in positions –1 to 6 are given, except for conserved hydrophobic residues in position 4 (marked by $), for both fingers of the proteins in subset A2. The helix positions are numbered relative to the beginning of the finger helix. A dot indicates an identical amino acid compared with the sequence of the first member of a subgroup. Key amino acid residues that are known or predicted to be essential for specific DNA base recognition are boxed and coloured, with Arg in red, His in blue, Asn in green and Asp in purple. However, we cannot exclude that some other amino acids (not boxed and coloured) in the given sequences also participate in base recognition.

from PROSITE (11). Other motifs, named BBOX (17), PCK-CRD (18), PHD and ZZ fingers (19,20), were taken from the literature and transformed into patterns containing only conserved zinc ligands with the spacings between them derived from the aligned sequences in the cited papers. Note that the program takes into account variants of BBOX (17) and RING (21) motifs with substituted zinc ligands in several positions and three additional motifs overlapping with the C2H2new pattern, named RPOL (RNA polymerase), C8 and NEW1 motifs (Fig. 2). The RPOL motif is a conserved zinc binding domain found in the N-terminal part of the largest subunit of RNA polymerases first described in Werner *et al*. (22). The C8 motif, currently formulated by us as C-$X_2$-C-$X_{10–22}$-C-$X_2$-C-$X_4$-C-$X_2$-C-$X_{10–17}$-C-$X_2$-C, matches members of three subtypes of Cys-rich motifs described recently in the zinc finger-like domains of the SSL1/BTF2 proteins (23), rabphilins (24) and the newly discovered zinc binding domain called the FYVE finger (25). The C8 motif resembles in its central part (underlined) the RING, LIM and PHD motifs. The NEW1 motif, in the form C-$X_2$-C-$X_9$-H-C-$X_2$-C-$X_2$-C-$X_5$-H-H-C-$X_5$-C, was derived from six ORFs of unknown function: four from the genome of *Saccharomyces* and one in each case from *Caenorhabditis* and *Schizosaccharomyces* (for more details on the BBOX variants and the C8 and NEW1 motifs consult the corresponding author: S.B., boehm@mdc-berlin.de).

As seen in Figure 2, 16 out of 22 'bad' fingers found in the yeast genome have clear overlaps with other motifs, namely to five RING, three GAL4, two GATA, two BBOX and in each case one ZZ, C8, RPOL and NEW1 motifs. Two 'bad' fingers have no overlaps with known motifs. Four out of the 22 wrong 'fingers' occur in yeast proteins which belong to well-known protein subfamilies (Fig. 2). These 'fingers' are not observed in other members of the particular subfamily and their occurrence in the four yeast proteins therefore is fortuitous. Taken together the C2H2new pattern detects 105 C2H2 zinc fingers in 53 proteins in

the yeast genome (Fig. 3), not taking into account the 22 wrong or questionable 'fingers'/proteins (Fig. 2).

### Classification of yeast C2H2 zinc finger proteins and functional implications

We have tried to classify the complete set of yeast Zfp in functional terms. It has been experimentally demonstrated that as a rule a tandem array comprising a minimum of two zinc fingers is required for sequence-specific high affinity DNA binding. In contrast, for most of the Zfp with a single finger or with dispersed fingers no sequence-specific DNA sites have been found so far. Therefore, we have classified the yeast Zfp according to the number, kind and arrangement (tandem arrays or dispersed) of their fingers into three subsets (A–C), as shown in Figure 3. All Zfp containing clusters with at least two fingers in tandem (linked by two to six residues) are put in subset A. Note that the subdivision of subset A into subgroups A1 (Zfp with tandem arrays of three or more fingers) and A2 (Zfp with a finger pair) is arbitrary, chosen here for discussion of sequence-specific DNA recognition of proteins with a finger pair (see below). Zfp with a single finger or with dispersed fingers (linked by 10 or more residues) are assigned to subset C. Zfp in subset B have finger arrangements in between subsets A and C and contain only one typical finger detected with C2H2new.

Experimental data have provided evidence that 10 out of the 30 Zfp included in subsets A1 and A2 (TF3A, RIM1, SWI5/ACE2, ADR1, MSN2/4, MIG1/2 and RGM) are involved in sequence-specific DNA binding (for references see yeast Zfp-related Internet resources). It can be predicted that tandem arrays of at least two or more fingers in the remaining 20 Zfp of subset A, which have not been characterized experimentally, recognize sequence-specific DNA sites as well. This prediction is based on application of the recently evolved finger/DNA recognition rules (5–8 and references therein) and on statistical analysis of sequence pattern conservation in a database of DNA recognition helices of fingers (S.Böhm *et al*., unpublished results). Here we will discuss in more detail only subset A2. This subset includes 22 proteins representing nearly half of all yeast Zfp. We wish to stress that seven out of the 22 proteins each contain one finger which is found with the C2H2new motif but not with C2H2ori. These additionally found fingers are predicted to be important for DNA binding specificity of the parent Zfp (see below).

The finger/DNA recognition rules relate the sequences of fingers to their preferred DNA binding sequences. These rules involve specific base contacts of particular amino acids in four key positions, –1, 2, 3 and 6, relative to the beginning of the finger helix (see the last two columns in subset A2 of Fig. 3). Exceptionally favourable is the presence of Arg in helix position –1, supported by Asp in position 2 and Arg in position 6, for recognition of guanine in GC-rich DNA stretches. This type of finger was found, for example, in the EGR and SP1 subfamilies of Zfp. Figure 3 shows the sequences of the DNA recognition helices in positions –1 to 6 for all yeast finger proteins with a finger pair. All 22 proteins (except two proteins of the NEW4 subgroup) have invariant Arg and Asp residues in helix positions –1 and 2 of their second finger and an invariant Arg in position 6 of their first finger, suggesting a central GG step in their DNA targets. Additional Arg residues are present in position –1 of finger 1 and/or in position 6 of finger 2 in the majority of the Zfp in subset A2. Moreover, most fingers of both types contain in position 3 further favourable residues, such as His or Asn, which recognize as a rule guanine or adenine respectively. These sequence features permit the classification of all mentioned yeast Zfp (except the two NEW4 proteins) in the group of EGR-like proteins which recognize purine (guanine)-rich DNA targets and which are known or predicted to be efficient gene regulatory proteins. Among them are all Zfp of the RGM1, NEW2 and NEW3 subgroups, each containing one finger detected only with the C2H2new motif.

Remarkably, all proteins in subset A2 (except HMS1) can be subdivided into eight subgroups, with two, three or five members respectively, based on significant sequence similarities in the finger clusters of their related members (data not shown). As a rule, the highest sequence conservation is found in the DNA recognition helices of the finger pairs belonging to the same subgroup (Fig. 3). Because of the conserved sequence pattern in the base recognition positions –1, 2, 3 and 6 of their finger helices the members of the different subgroups are predicted to bind to identical or very similar DNA sequences. The same holds true for the Zfp of the SWI5 subgroup, with a highly conserved finger triplet. Indeed, recent experimental data for the homologous Zfp pairs SWI5 and ACE2 (29), MSN2 and MSN4 (30) and MIG1 and MIG2 (31) are in agreement with this prediction but also highlight functional differences between the mentioned protein pairs arising from non-homologous sequences outside their finger cluster.

Interestingly, our results also show that, in contrast to higher eukaryotes, the yeast genome does not contain multifinger proteins with 10 or more repeats (exemplified for example by Zfp with a KRAB domain). Only three proteins (TF3A, YJL056 and AZF1; see subset A1 in Fig. 3) contain tandem arrays of four or more fingers. It is tempting to speculate that the lack of multifinger proteins correlates rather well with the compactness of the yeast genome (12), with short intergenic regions and a scarcity of introns and repeated sequences contrasting greatly with the genomes of higher eukaryotes.

## ACKNOWLEDGEMENTS

We would like to thank G.H.Jacobs for providing his Zf/Zfp database, P.S.Freemont and R.Aasland for help in classification of RING, PHD and FYVE fingers respectively and U.Heinemann for improving the manuscript. S.B. was supported by the Deutsche Forschungsgemeinschaft through the SFB 344/Projekt YE3.

## REFERENCES

1 Ginsberg,A.M., King,B.O. and Roeder,R.G. (1984) *Cell*, **39**, 479–489.
2 Miller,J., McLachlan,A.D. and Klug,A. (1985) *EMBO J.*, **4**, 1609–1614.
3 Böhm,S. and Drescher,B. (1985) *Studia Biophys.*, **107**, 237–247.
4 Brown,R.S., Sander,C. and Argos,P. (1985) *FEBS Lett.*, **186**, 271–274.
5 Klug,A. and Schwabe,J.W.R. (1995) *FASEB J.*, **9**, 597–604.
6 Kim,C.A. and Berg,J.M. (1996) *Nature Struct. Biol.*, **3**, 940–945.
7 Greisman,H.A. and Pabo,C.O. (1997) *Science*, **275**, 657–661.
8 Choo,Y. and Klug,A. (1997) *Curr. Opin. Struct. Biol.*, **7**, 117–125.
9 Hoovers,J.M.N., Mannens,M., John,R., Blick,J., van Heyningen,V., Porteus,D.J., Leschot,N.J., Westerveld,A. and Little,P.F.R. (1992) *Genomics*, **12**, 254–2637.
10 Jacobs,G.H. (1992) *EMBO J.*, **11**, 4507–4517.
11 Bairoch,A., Bucher,P. and Hofmann,K. (1995) *Nucleic Acids Res.*, **24**, 189–196.
12 Goffeau,A. *et al*. (1996) *Science*, **274**, 562–567.
13 Schwabe,J.W.R. and Klug,A. (1994) *Nature Struct. Biol.*, **1**, 345–349.

14 Thompson,J.D.,Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**,4673–4680.
15 Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.
16 Kuhner,M.K. and Felsenstein,J. (1994) *Mol. Biol. Evol.*, **11**, 459–468.
17 Reddy,B.A., Etkin,L.D. and Freemont,P.S. (1992) *Trends Biochem. Sci.*, **17**, 344–345.
18 Hommel,U., Zurini,M. and Luyten,M. (1994) *Nature Struct. Biol.*, **1**, 384–388.
19 Aasland,R., Gibson,T.J. and Stewart,A.F. (1995) *Trends Biochem. Sci.*, **20**, 56–59.
20 Ponting,C.P., Blake,D.J., Davies,K.E., Kendrickk-Jones,J. and Winder,S.J. (1996) *Trends Biochem. Sci.*, **21**, 11–13.
21 Borden,K.L.B. and Freemont,P.S. (1996) *Curr. Opin. Struct. Biol.*, **6**, 395–401.
22 Werner,M., Denmat,S.H., Treich,I., Sentenac,A. and Thuriaux,P. (1992) *Mol. Cell. Biol.*, **12**, 1087–1095.
23 Humbert,S., van Vuuren,H., Lutz,Y., Hoeijmakers,H.J., Egly,J.-M. and Moncollin,V. (1994) *EMBO J.*, **13**, 2393–2398.
24 Stahl,B., Chou,J.H., Li,C., Südhof,T.C. and Jahn,R. (1996) *EMBO J.*, **15**, 1799–1809.
25 Stenmark,H., Aasland,R., Toh,B.H. and D'Arrigo,A. (1996) *J. Biol. Chem.*, **271**, 24048–240541.
26 McIntosh,E.M. and Haynes,R.H. (1986) *Mol. Cell. Biol.*, **6**, 1711–1721.
27 Krisak,L., Strich,R., Winter,R.S., Hall,J.P., Mallory,M.J., Kreitzer,D., Tuan,R.S. and Winter,E. (1994) *Genes Dev.*, **8**, 2151–2161.
28 Bartel,B., Wuenning,I. and Varshavsky,A. (1990) *EMBO J.*, **9**, 3179–3189.
29 Dohrmann,P.R., Voth,W.P. and Stillman,D.J. (1996) *Mol. Cell. Biol.*, **16**, 1746–1758.
30 Martinez-Pastor,M.T., Marchler,G., Schuller,C., Marchlar-Bauer,A., Ruis,H. and Estruch,F. (1996) *EMBO J.*, **15**, 2227–2235.
31 Lutfiyya,L.L. and Johnston,M. (1996) *Mol. Cell. Biol.*, **16**, 4790–4797.