SURVEY AND SUMMARY

Cytoplasmic ribosomal protein genes of the fission yeast Schizosaccharomyces pombe display a unique promoter type: a suggestion for nomenclature of cytoplasmic ribosomal proteins in databases

Thomas Gross and Norbert F. Käufer*

Institut für Genetik-Biozentrum, Technical University of Braunschweig, Spielmannstrasse 7, 38106 Braunschweig, Germany

Received March 17, 1998; Revised and Accepted May 18, 1998 DDBJ/EMBL/GenBank accession nos AJ002733 and AJ002734

ABSTRACT

We identified 34 new ribosomal protein genes in the Schizosaccharomyces pombe database at the Sanger Centre coding for 30 different ribosomal proteins. All contain the Homol D-box in their promoter. We have shown that Homol D is, in this promoter type, the TATA-analogue. Many promoters contain the Homol E-box, which serves as a proximal activation sequence. Furthermore, comparative sequence analysis revealed a ribosomal protein gene encoding a protein which is the equivalent of the mammalian ribosomal protein L28. The budding yeast Saccharomyces cerevisiae has no L28 equivalent. Over the past 10 years we have isolated and characterized nine ribosomal protein (rp) genes from the fission yeast S.pombe. This endeavor yielded promoters which we have used to investigate the regulation of rp genes. Since eukaryotic ribosomal proteins are remarkably conserved and several rp genes of the budding yeast S.cerevisiae were sequenced in 1985, we probed DNA fragments encoding S.cerevisiae ribosomal proteins with genomic libraries of S.pombe. The deduced amino acid sequence of the different isolated rp genes of fission yeast share between 65 and 85% identical amino acids with their counterparts of budding yeast.

ARCHITECTURE OF PROMOTER REGIONS OF RIBOSOMAL PROTEIN GENES

Molecular dissection of ribosomal protein (rp) gene promoter regions of Schizosaccharomyces pombe revealed a promoter type which does not contain a canonical TATA-box. Instead, these promoters display a TATA-analogue named the Homol D-box. We showed that the Homol D-box, represented by the sequence CAGTCACA or its reverse complementary sequence TGTGAC-TG, is involved in determining transcriptional start sites and is the target of protein factor(s) binding. The binding of this factor cannot be competed with TATA-box containing oligonucleotides (1). The Homol D-box has been compared with the TATA-box with respect to its potential to form local sequence-specific structures which may contribute to the binding specificity of trans-acting proteins (2; Table 1). In some rp promoters we found the tandem repeat AGGG-TAGGGT or its reverse complementary sequence ACCCTACCCCT upstream of the Homol D-box. We named this sequence the Homol E-box and showed that it functions, in the proximal arrangement, with Homol D as an activation sequence. It is also the target of protein factors (3; Table 1).

The Sanger Centre (Cambridge, UK) together with 12 other European laboratories, is sequencing the S.pombe genome (http://www.sanger.ac.uk/Projects/S_pombe/). Finished and unfinished sequences are available and searchable. We searched these sequences, represented by cosmids of chromosomes I and II, and found 43 genes coding for ribosomal proteins, including several we and others had already sequenced (Table 1).

Strikingly, a comparison of the promoter region of the rp genes reveals a statistically significant pattern for Homol D- and Homol E-boxes. First, all promoters contain the Homol D-box in a position between 49 and 102 bp upstream of the start codon ATG (Table 1). Second, those promoters which unequivocally contain, in addition to Homol D, the Homol E-box, display Homol E as a proximal upstream element of Homol D. The distance between the two elements ranges from 0 to 14 bp with no obvious bias for a preferred distance (Table 1). It is also noticeable from Table 1 that both sequences, Homol D and Homol E, are highly conserved. Only one sequence of Homol D and one of Homol E show two changes in the canonical Homol-box (rpaP1-1, rp126, Table 1). Fewer than 8% show a change of one nucleotide. For Homol D we found CAGTCACG instead of CAGTCACA, and AGTGACTG instead of TGTGACTG. The tandem repeat AGGG-TAGGGT of Homol E was found as AGGGTAGGGC or as AGGTTAGGGT, whereas the reverse complementary repeat ACCCTACCCCT was always found as GCCCTACCCCT or ACCCTAACCCT. In the Homol E sequence, two nucleotides were changed as indicated in Table 1 (rp126; Table 1).

*To whom correspondence should be addressed. Tel: +49 531 391 5774; Fax: +49 531 391 5765; Email: n.kaeuer@tu-bs.de
We compared the promoter strength of two gene families (rp127a and rp18; Table 1) using lacZ as a reporter. One family consisted of three genes with two promoters displaying only Homol D and one containing the Homol E/Homol D arrangement; the other family had a promoter with a Homol D-box and a promoter with the Homol E/Homol D arrangement. These two families express equal amounts of β-Gal activity (3). However, these data do not reveal whether Homol E plays a role in the regulation of rp genes. The data show that Homol E is a proximal transcription activation sequence which can activate through the Homol D-box (3). Moving Homol E more than 21 bp away from Homol D leads to a drastic decrease in β-Gal activity, indicating that it is a proximal promoter element (I. Witt, personal communication).

The compilation of the rp gene promoters (Table 1) confirms our data and predictions published previously (1, 3). We suggested that rp genes of fission yeast display a promoter type in which Homol D mediates basal transcription and Homol E is a proximal activation sequence. Here we propose that Homol E and Homol D, their relative position to each other and the position of Homol D relative to the start codon are used as signatures for this promoter type.

### INTRONS IN RP GENES

The architecture of the promoter of rps5a-1 appears to be an exception to the rule since the distance between the Homol D-box and the start codon ATG is 213 bp (Table 1). Interestingly, comparative analysis of this promoter region reveals 5' and 3' splice sites including a branch site establishment possible S. pombe introns of 147 or 87 bp, respectively, in this untranslated region (5). We found introns in 40% of the 51 rp genes. A random count of 200 ORFs in the S. pombe database counting intron-less versus intron containing ORFs revealed 45% intron containing genes. This is consistent with the results of a survey which we conducted a few years ago. We searched a random sample of 100 genes for introns and found that 40% of them contained introns (5). Thus rp genes contain introns with a frequency found for all the other genes.

The distribution of the introns in rp genes seems to be random. The smallest exon1 found in an ORF consists of just the sequence. Here we propose that Homol E and Homol D, their proximal position to each other and the position of Homol D relative to the start codon are used as signatures for this promoter type.

### SUGGESTION FOR NOMENCLATURE OF S. POMBE RP GENES AND PROTEINS

The sequence of the genome of the budding yeast S. cerevisiae has been completed and is available for comparative analysis. Based on such a comparative analysis a new nomenclature for ribosomal proteins of S. cerevisiae has been proposed (7). The new numbering system of ribosomal proteins of S. cerevisiae follows, when possible, the nomenclature for mammalian ribosomal proteins. Significant differences still exist. One major difference
is that there is no *S.cerevisiae* counterpart for mammalian ribosomal protein L28 (7,8).

We compared the amino acid sequences of the *S.pombe* ORFs identified in the Sanger Centre (Table 1) with the available databanks using BLAST 2.0, a similarity search tool developed by Altschul *et al.* (9). With this approach we found a gene on cosmid C2E11 (Table 1) which aligns with sequences from rat, *Caenorhabditis elegans* and *Trypanosoma cruzi*. Notably, the rat sequence (Fig. 1A, rnL28) shares 37% identical amino acids with the *S.pombe* sequence (Fig. 1A, spL28), and 36 and 31% with *C.elegans* and *Trypanosoma*, respectively. The rat sequence is described in the database as ribosomal protein L28 (Fig. 1A, rnL28). The ribosomal proteins of rat are the base for the proposed mammalian numbering system since the sequence determination of all rat ribosomal proteins has been almost completed (8; Table 1).

We did not find an ORF in the genomic sequence of *S.cerevisiae* bearing significant similarity to these sequences. All the other ORFS of *S.pombe* resemble ribosomal protein sequences of *S.cerevisiae* and rat, whereas in general the sequences share between 60 and 85% identical amino acids. For example, the ORF on cosmid C23C11 (Table 1) aligns with ribosomal protein S23 of *S.cerevisiae* (scS23) and rat (rnS23), sharing 85 and 75% identical amino acids, respectively (Fig. 1B). According to the new *S.cerevisiae* nomenclature, this rp has recently been designated S23 (7). Furthermore, the ORF on cosmid C19G10 (Table 1) aligns with ribosomal protein L23a of rat (rnL23a), sharing 64 and 61% identical amino acids, respectively (Fig. 1C). L25 is the designation of the *S.cerevisiae* rp following the new nomenclature (7, Table 1). These two examples reflect the range of similarity found between the species for all the sequences which we have tested, except for L28 (Fig. 1).

We also found rp genes in the *S.pombe* databank encoding the counterparts of acidic ribosomal proteins (10). Based on sequences taken from the literature (10), we identified two promoters of acidic rp genes as containing canonical TA TA-boxes (1). One of these genes (*rpa3*), was found in our search of the *S.pombe* database for promoters containing Homol E-/Homol D-boxes (Table 1). The two promoter regions are completely different in sequence, whereas the sequences spanning the structural gene, the intron sequences and the 3′ trailer of the gene are identical. This indicates that we identified the authentic *rpa3* gene. The TA T A-box promoter in front of *rpa3* must be a cloning accident (10).

Based on the data presented here, we suggest the *S.pombe* genome project and the scientific community use the mammalian nomenclature for the cytoplasmic ribosomal proteins (7,8).
also suggest indicating the fission yeast ribosomal protein genes as \textit{rps3-1, rps3-2, rp17a-1, rp130}, etc. If there are two copies, they should be labeled -1, -2, etc. (Table 1). The proteins coded by the genes should be called Rps3.1, Rps3.2, Rp17a.1, Rp130, etc. (Table 1). The gene names \textit{rps} and \textit{rp1} for ribosomal proteins of the small and large subunit, respectively, have been proposed by Kohli (11). The acidic ribosomal protein genes isolated by Beltrame and Bianchi (10) have been named \textit{rpa1, rpa2, rpa3} and \textit{rpa4}. In the new \textit{S.cerevisiae} numbering system and in the mammalian system (7), the acidic equivalents of Rpa1 and Rpa3 are designated P1; the equivalents of Rpa2 and Rpa4 are named P2. In addition, we found in the EMBL databank two cDNAs named \textit{rpa5} and \textit{rpa6} (AJ002733 and AJ002734). Rpa5 belongs to the P1 and Rpa6 belongs to the P2 family of acidic ribosomal proteins. Therefore, we suggest the designation RpaP1.1, RpaP1.3 and RpaP1.5, and RpaP2.2, RpaP2.4 and RpaP2.6 for these two families (Table 1).

Since many ribosomal protein genes are represented by two copies and 35 different ribosomal proteins have been identified in the \textit{S.pombe} genome project, we would expect that at least another 50 ribosomal protein genes will be sequenced and named. If the mammalian nomenclature is used and the new nomenclature of the \textit{S.cerevisiae} counterparts is always added in brackets (7), we are confident that the currently unbearable confusion in nomenclature of ribosomal proteins will be under control.

**ACKNOWLEDGEMENT**

We thank Drs Isabell Witt and Anita Marchfelder for communicating unpublished results.

**REFERENCES**