J Mol Evol (2007) 65:529–540
DOI 10.1007/s00239-007-9031-y

Genomic Evolution of the Proteasome System Among Hemiascomycetous Yeasts

Gertrud Mannhaupt · Horst Feldmann

Received: 14 April 2007 / Accepted: 17 August 2007 / Published online: 2 October 2007

Abstract Components of the proteasome-ubiquitin pathway are highly conserved throughout eukaryotic organisms. In *S. cerevisiae*, the expression of proteasomal genes is subject to concerted control by a transcriptional regulator, Rpn4p, interacting with a highly conserved *cis*-regulatory element, PACE, located in the upstream regions of these genes. Taking advantage of sequence data accumulated from 15 *Hemiascomycetes*, we performed an in silico study to address the problem of how this system might have evolved among these species. We found that in all these species the Rpn4p homologues are well conserved in terms of sequence and characteristic domain features. The "PACE patterns" turned out to be nearly identical among the *Saccharomyces* "sensu stricto" species, whereas in the evolutionary more distant species the putatively functional *cis*-regulatory motifs revealed deviations from the "canonical" PACE nonamere sequence in one or two nucleotides. Our findings suggest that during evolution of the *Hemiascomycetes* such slightly divergent ancestral motifs have converged into a unique PACE element for the majority of the proteasomal genes within the most recent species of this class. Likewise, the Rpn4 factors within the most recent species of this class show a higher degree of similarity in sequence than their ancestral counterparts. By contrast, we did not detect PACE-like motifs among the proteasomal genes in other eukaryotes, such as *S. pombe*, several filamentous fungi, *A. thaliana*, or humans, leaving the interesting question which type of concerted regulation of the proteasome system has developed in species other than the *Hemiascomycetes.*

Keywords *Hemiascomycetes* · In silico analysis · Proteasome · Regulation · Yeast

Electronic supplementary material The online version of this article (doi:10.1007/s00239-007-9031-y) contains supplementary material, which is available to authorized users.

G. Mannhaupt
Institute for Bioinformatics, GSF, Ingolstaedter Landstr. 1, D-85764 Neuherberg, Germany

H. Feldmann
Molecular Biology, Adolf-Butenandt-Institute, Schillerstr. 44, D-80336 München, Germany

H. Feldmann (✉)
Ludwig-Thoma-Str. 22B, D-85232 Bergkirchen, Germany
e-mail: horst.feldmann@med.uni-muenchen.de

Introduction

In recent years, a variety of investigations has been conducted aimed at identifying DNA binding proteins and conserved transcription factor binding sites genome-wide in a single species or in evolutionarily related species. The genome of *Saccharomyces cerevisiae* (e.g., Gasch et al. 2000; Hughes et al. 2000) and genomes of the genus *Saccharomyces* (e.g., Kellis et al. 2003; Chiang et al. 2003; Cliften et al. 2003; Moses et al. 2003, 2004) were used as convenient models in these studies. Recently, Gasch et al. (2004) have extended the analysis to 10 *Hemiascomycete* species and four other *Ascomycete* species.

In the year 2000, a couple of French laboratories had started the Génolevures project (Souciet et al. 2000), with the goal of obtaining genomic information from 13 species of the class *Hemiascomycetes*, simple fungi the vast majority of which are yeasts, that could be used to study the level of genetic diversity between these yeast species and the level of protein divergence within them (Malpertuy et al. 2000) as well as the level of synteny conservation

between these genomes (Llorente et al. 2000; Fischer et al. 2001). These studies established that the level of genetic diversity between yeast species is often unsuspected. For instance, the average protein divergence of >50% found between *Saccharomyces cerevisiae* and *Yarrowia lipolytica* revealed that *Hemiascomycetes* are molecularly as diverse as the entire phylum of chordates (Makalowski et al. 1998; Dujon 2006).

Meanwhile, 15 more or less complete annotated genome sequences are available from the *Hemiascomyctes* class (see below). This information can be used to exploit the genomic evolution of regulatory networks in these species, including the genes regulated by specific transcription factors and the cognate *cis*-regulatory elements interacting with these factors.

In the present study we have focused on the 26S proteasome system, as it meets the requirements for a thorough in silico analysis. The 26S proteasome, responsible for the programmed proteolysis of proteins, has been intensely studied in *S. cerevisiae*, and comparisons with higher eukaryotic organisms have shown that at least the components of the 20S core particle and the six AAA^+ ATP-binding proteins (RPTs) of the 19S cap particle are highly conserved from yeast to mammals (for an overview see Wolf and Hilt 2004). The *S. cerevisiae* genes for these entities are essential and single-copy throughout. Nearly all of the genes encoding the subunits of the 20S core (except *PRE5*) and the 19S regulatory particle (except *RPN8*, *RPN10*, and *RPN13*) possess a unique (nondegenerate) upstream nonamer box (GGTGGCAAA) which we called PACE and which was shown to bind to Rpn4p, a transcriptional activator (Mannhaupt et al. 1999). Further studies have elucidated that Rpn4p is a ligand, substrate, and transcriptional regulator of the 26S proteasome and exerts a negative feedback control (e.g., Xie and Varshavsky, 2001; Ju and Xie, 2004; Wang et al. 2004) and that Rpn4p participates in regulatory networks such as DNA damage repair (Jelinsky et al. 2000), stress responses (Owsianik et al. 2002), and filamentous growth (Prinz et al. 2004).

We compared the relevant elements of the proteasome system from *Hemiascomycetes* using those from *S. cerevisiae* as a reference genome. Our study revealed that these elements are highly conserved among the *Hemiascomycetes*, suggesting that similar control mechanisms of the proteasome system are operative among these yeast species. Extending the comparisons to data from *S. pombe*, three filamentous fungi, *A. thaliana*, and human, we did not detect true counterparts for Rpn4p or PACE-like elements in the latter species. These observations suggest that the regulation of the proteasome system in species other than the *Hemiascomycetes* is subject to different, but still unknown, control mechanisms.

Methods

Retrieval and Comparison of Gene Sequences

Orthologues for the 20S core subunits, the 6 RPT subunits, and the 14 RPN subunits of the 19S cap as well as those for Uba1p and Cdc48p were retrieved by searching the MIPS PEDANT databases (http://www.pedant.gsf.de) with the BLAST algorithm (Altschul et al. 1990). Data collections and references for the original genome sequences are as follows: http://www.yeastgenome.org or http://www.mips.gsf.de/genre/proj/yeast/ for *S. cerevisiae* (Goffeau et al. 1996); http://www.broad.mit.edu/annotation/fungi/comp_yeasts/ for *S. paradoxus, S. mikatae*, and *S. bayanus* (Kellis et al. 2003); http://www.genome.wustl.edu/ for *S. castellii*, *S. kluyveri*, and *S. kudriavzevii* (Cliften et al. 2003); http://www.broad.mit.edu/ for *K. waltii* (Kellis et al. 2004); http://www.cbi.labri.fr/Genolevures/ for *C. glabrata, K. lactis, D. hansenii*, and *Y. lipolytica*; http://www.agd.unibas.ch for *A. gossypii* (Dietrich et al. 2004); http://www.candidagenome.org for *C. albicans* (Jones et al. 2004) and *C. dubliniensis*; http://www.genedb.org/genedb/pombe/ for *S. pombe* (Wood et al. 2002); database *N. crassa* (Galagan et al. 2003) and http://www.mips.gsf.de/genre/proj/ncrassa/ for *N. crassa*; http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans/Home.html for *A. Nidulans*; Aspergillus fumigatus genome project http://www.sanger.ac.uk/Projects/A_fumigatus/ for *A. fumigatus* (Nierman et al. 2005); http://www.mips.gsf.de/proj/plant/jsf/athal/index.jsp for *A. thaliana* (Arabidopsis Genome Initiative 2000); and Human Genome Resources http://www.ncbi.nlm.nih.gov/genome/guide/human/ for *H. sapiens*.

Searches for PACE-like Upstream Sequences

Five hundred base pairs of 5′-upstream sequence for each *S. cerevisiae* gene was extracted from a file provided on the MIPS FTP server (ftp://www.ftpmips.gsf.de/yeast/sequences/Scerevisiae_utr5_500.fa). For all other species, 500 bp of 5′-upstream sequences was extracted from the respective PEDANT database by internal accession using appropriate MYSQL queries. Multiple FASTA files containing these species specific promoter sequences were used as input into a JAVA based pattern search program, listing all patterns of 9mers, occurring on all promoters on both strands. In a second step, the PACE motif known from *S. cerevisiae* (GGTGGCAAA) was used as a search pattern, allowing two mismatches. The output for each species lists the resulting motifs in descending frequency and their positions together with the codes for the respective proteins.

The RSA tool (http://www.rsat.ulb.ac.be/rsat/) was used to list PACE or PACE-like sequences in the *Hemiascomycetes* species included in this program.

Alignment Tools

Alignment of the Rpn4p orthologues or upstream promoter sequences of proteasomal genes was done using the CLUSTAL W routine at the EBI server (http://www.ch.embnet.org) or DiAlign (http://www.dialign.gobics.de [Morgenstern et al. 2006]).

Results

Similarities of Proteasomal Gene Products from Other Species to *S. cerevisiae*

The sequences for 12 gene products of the 20S core, the 6 RPTs, and the *RPN* gene products from the 19S cap particle as well as those of the homologues of Uba1p and Cdc48p from the 15 *Hemiascomyctes* species and the 6 "outgroup" species (*S. pombe*, *N. crassa*, *A. nidulans*, *A.fumigatus*, *A. thaliana*, and *H. sapiens*) analyzed here were retrieved and compared as described under Methods. The sequences from *S. cerevisiae* were taken as a reference; we felt that pairwise comparison of all of these sequences was unnecessary. The results are presented in Table 1. For simplicity of discussion, we have divided the 15 *Hemiascomycetes* species into three groups: group 1 comprises *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, and *S. kudriavzevii* (whereby this group represents the *Saccharomyces* "sensu strictu" species); group 2 comprises *S. castelli, S. kluyveri*, *A. gossypii*, *K. lactis*, *K. waltii*, and *C. glabrata*; and group 3 comprises *C. albicans*, *C. dubliniensis, D. hansenii*, and *Y. lipolytica*.

The conservation of the 20S core proteolytic subunits (Table 1A) is remarkably high: it ranges from 98% to 100% similarity within group 1 and from 81% to 97% within group 2. A noticeable decrease in similarity is seen in *C. albicans* and *C. dubliniensis* as well as in *D. hansenii* and *Y. lipolytica* (group 3), from 67% to 88%. Similarities within the "outgroup" species are still remarkably high (61% to 86%) but below those within group 3.

Sequence similarities among the Rpt products (Table 1 A) are even higher than those observed for the 20S core subunits: 98%–100% within the *Saccharomyces* species (group 1) and 89%–99% in group 2. A slight decrease in similarity is seen for *C. albicans* and *C. dubliniensis* (85%–94%) as well as for *D. hansenii* (84%–94%) and *Y. lipolytica* (86%–94%). Remarkably, there is still high similarity for the "outgroup" species (81%–89%). An interesting finding was that in *Arabidopsis* many proteasomal genes (as far as sequences were available) are duplicated. It is noteworthy to stress that in none of the *Hemiascomycetes* species or in the other species were duplicates for any of the proteasomal genes detected. For the species listed in the *Yeast Genome Order Browser* (YGOB; http://www.wolfe.gen.tcd.ie/ygob) (Scannel et al. 2006), this was verified by looking up all genes relevant for our study. Interestingly, *S. castellii* has a second gene product, each with similarity to Rpn4 and Cdc48, respectively. However, the second copy of *S. castellii* Rpn4 (713.11) might be a nonfunctional relic, as similarities at the N-terminus and within the regions where the acidic domains are located are largely lost. Therefore we have relied in comparisons only on the first copy of Rpn4p (718.67), which has retained the characteristic features throughout the sequence.

Compared to the Rpt and 20S moieties, the *RPN* gene products on average are less conserved with reference to those of *S. cerevisiae* (Table 1B). This may be due to the fact that the single species have different lifestyles, and hence the functions of the Rpn proteins had to be adapted correspondingly. Note, for example, that subunits similar to Rpn13p and/or Rpn14p may be even missing from some species.

The most pronounced deviations in similarity become apparent when comparing the Rpn4p homologues: while the similarity ranges from 86% to 92% in the *Saccharomyces* "sensu stricto" (group 1), there is a sudden drop in similarity (39%–50%) when the Rpn4p homologues of the remaining *Hemiascomycetes* species are compared to Rpn4p from *S. cerevisiae* (Table 1B), mainly due to substantial variations within the central part. Therefore, we have aligned all Rpn4p-like sequences retrieved from the databases and carefully checked them for the presence of their known characteristic features, the highly conserved atypical Zn-finger at the C-terminus, and the two acidic domains in the center of the sequence as observed in *S. cerevisiae* (Mannhaupt et al. 1999). These features were found to be highly conserved among the Rpn4p homologues from the group 1 species (Fig. 1). Among the remaining *Hemiascomycete* species (groups 2 and 3), the Rpn4p homologues exhibit the conserved atypical Zn-finger at the C-terminus, which is always the most highly conserved part of the sequence, because it represents the DNA-binding domain (see also Gasch et al. 2004). Though deviating in sequence and length, two acidic domains are present in the Rpn4p homologues of all *Hemiascomycetes* species, occurring in similar relative positions (Fig. 1 and Supplement 1). Therefore, we conclude that the Rpn4p homologues from all *Hemiascomycetes* represent true transcription factors involved in the regulation of the proteasomal and further genes in these organisms. By contrast, the only conserved feature in the Rpn4p-like sequences from the "outgroup" species is the occurrence of the highly conserved C-terminal Zn-finger (see Supplement 2). Note, however, that the loops between CxxC and HxxxxH of the Zn-finger are shorter (14 instead of 21) in these cases than

Table 1 Homologies of 20S and 19S components from various species vs. *S. cerevisiae*

(A) 20S proteolytic, 19S ATPase subunits and Uba1p/Cdc48p

	PRE1	PRE2	PRE3	PRE4	PRE5	PRE6	PRE7	PRE8	PRE9	PRE10	PUP2	PUP3	Average PRE1-PUP3	RPT1	RPT2	RPT3	RPT4	RPT5	RPT6	Average RPT1-6	Uba1	Cdc48
S. paradoxus	99	99	99	98	99	99	99	99	99	99	97	99	98.8	99	100	99	97	100	100	99.2	98	99
S. mikatae	98	na	99	98	98	99	99	99	99	98	99	na	98.6	99	100	100	97	99	100	99.2	na	98
S. bayanus	98	98	98	98	98	99	98	99	99	98	98	99	98.3	99	100	99	97	99	100	99.0	97	98
S. kudriavzevii	98	98	98	97	98	99	98	99	na	98	99	99	98.4	100	na	98	98	99	100	99.0	na	99
S. castellii	96	92	87	83	89	96	96	94	97	96	95	97	93.2	93	99	96	93	97	97	95.8	92	92 93
S. kluyveri	94	90	na	85	91	na	95	92	96	82	na	96	91.2	94	98	96	95	95	97	95.8	na	92
C. glabrata	96	89	92	85	92	91	96	96	94	90	94	98	92.8	91	99	91	93	96	98	94.7	91	95
K. waltii	93	92	95	83	88	92	94	94	90	84	94	95	91.2	91	95	90	89	95	95	92.5	88	94
K. lactis	90	92	95	83	89	94	91	93	91	90	90	96	91.2	90	98	93	89	94	96	93.5	89	91
A. gossypii	92	87	93	81	88	93	90	92	92	81	93	95	89.8	92	96	94	90	95	97	94.0	86	92
C. albicans	85	85	86	70	76	85	88	80	82	76	88	84	82.0	85	94	91	86	89	91	89.0	83	90
C. dubliniensis	85	86	86	71	77	na	87	80	83	77	88	na	82.0	86	91	91	87	89	93	89.5	83	90
D. hansenii	84	86	87	72	77	86	88	84	80	80	89	87	83.3	86	94	87	84	89	93	88.8	84	91
Y. lipolytica	82	85	88	67	77	84	81	89	82	81	88	86	82.5	86	93	94	86	89	90	89.0	78	88
N. crassa	76	76	84	67	70	84	72	84	79	75	84	86	78.1	84	84	85	85	88	89	85.8	74	85
A. fumigatus	78	77	86	63	72	82	70	75	78	70	81	82	76.2	84	84	90	85	88	88	86.5	77	88
A. nidulans	75	78	86	62	72	83	73	76	78	73	83	83	76.8	83	84	86	85	87	88	85.7	76	85
S. pombe	73	80	81	64	77	75	76	80	79	70	82	80	76.4	82	85	87	86	83	89	85.3	73	86
H. sapiens	65	80	76	65	69	78	64	67	73	68	74	75	71.2	82	85	81	83	86	89	84.3	71	85
A. thaliana (a)	59	76	73	62	65	76	61	76	77	66	76	71	69.8	81	85	85	82	85	86	84.0	63	83
A. thaliana (b)	59	75			65	77		76			76	72		78	85	Na	80	85	85		65	82

(B) 19S non-ATPase subunits

	RPN1	RPN2	RPN3	RPN5	RPN6	RPN7	RPN8	RPN9	RPN10	RPN11	RPN12	Average1–12	RPN13	RPN14	RPN4
S. paradoxus	98	99	98	99	100	98	98	99	97	99	98	99.4	96	91	92
S. mikatae	97	97	95	94	98	98	98	97	na	96	96	96.5	94	89	90
S. bayanus	96	97	95	98	96	97	96	97	96	99	95	96.5	94	82	86
S. kudriavzevii	94	96	95	97	pt 99	pt. 90	97	96	96	?na	95	95.4	90	84	90
S. castellii	86	85	79	90	pt 96	84	89	80	85	95	77	83.8	81	62	50 46
S. kluyveri	84	85	76	87	86	84	88	83	85	92	75	84.1	72	59	na
C. glabrata	85	86	79	89	91	85	90	78	81	95	77	83.3	72	57	49
K. waltii	82	85	76	87	85	83	83	85	86	89	75	83.3	74	56	47
K. lactis	81	81	75	88	85	82	88	79	82	90	69	81.8	?	61	46

Table 1 continued

(B) 19S non-ATPase subunits

	RPN1	RPN2	RPN3	RPN5	RPN6	RPN7	RPN8	RPN9	RPN10	RPN11	RPN12	Average1–12	RPN13	RPN14	RPN4
A. gossypii	83	80	76	86	85	83	85	79	87	90	70	82.2	68	59	42
C. albicans	70	76	60	71	72	70	78	57	71	85	57	69.7	77	50	39
C. dubliniensis	69	74	61	66	73	71	77	59	73	84	57	69.5	59	na	37
D. hansenii	70	73	61	69	71	70	77	63	69	85	60	70.2	62	48	40
Y. lipolytica	68	69	58	75	69	76	77	61	71	81	54	69.0	52	na	39
N. crassa	63	59	56	70	66	44	75	59	68	80	51	62.8	51	50	(44)
A. fumigatus	64	64	58	63	65	54	76	52	65	79	51	62.8	na	na	(39)
A. nidulans	63	55	59	66	64	54	73	56	64	82	51	62.5	na	na	(46)
S. pombe	57	61	58	64	66	61	72	58	64	76	53	62.7	pt 49	pt 41	(38)
H. sapiens	58	59	56	65	65	58	69	58	66	80	na	61.9	47	47	(38)
A. thaliana (a)	70	63	57	60	67	62	73	57	68	76	51	63.6	Na	na	na
A. thaliana (b)	57	64	58	60			74	56							

Note. The numbers refer to percentage similarity in protein sequence. na, sequence not available; pt, only part of sequence available. The following functions have been ascribed to particular Rpn proteins from *S. cerevisiae:* Rpn1, ligand binding; Rpn 2, binding of ubiquitin ligase Hul5 ; Rpn3, cell cycle control; Rpn5-7; PCI-domain lid subunits, Rpn8, MPN domain protein; Rpn9, cell cycle control and assembly of proteasome; Rpn10, poly-ubiquitin binding; Rpn11, metalloprotease-like deubiquitinating activity; Rpn12, interaction with Rpn3

for the rest of the sequences, and that the protein sequence from *H. sapiens* is even considerably shorter. In none of the "outgroup" Rpn4p-like sequences could we detect any acidic domains.

Presence of PACE-like Sequences in *S. cerevisiae* and Other Species

Next we inspected the 5′-upstream sequences of the proteasomal genes from the *Hemiascomycetes* as well as those of *UBA1* (ubiquitin activating enzyme; E1) and *CDC48* (ATPase in ER, nuclear membrane, and cytosol) for the occurrence of PACE or PACE-like elements. Both are single-copy genes in *S. cerevisiae* and belong to a large group of genes that appear to be under the control of Rpn4p (Mannhaupt et al. 1999; Kapranov et al. 2001). Interestingly, the PACE box is fully conserved in the majority of the *CDC48* promoters (see Table 2), except in *D. hansenii* and *Y. lipolytica*. As indicated in Table 1A, the homologues of Uba1p and Cdc48p on average share an even higher degree of similarity with their counterparts from *S. cerevisiae* than the proteasomal genes, pointing to their absolute requirement throughout eukaryotes.

An interval of 500 bp upstream from the translational start site was chosen, as the elements in *S. cerevisiae* are located within this region, varying between position –83 and position –163. Likewise, searches for PACE or PACE-like elements in *Hemiascomycetes* as far as they are available for the RSA (regulatory sequence analysis) tools (e.g., van Helden 2003) indicated their presence upstream of proteasomal gene promoters in noncoding regions. Our JAVA program (see Methods) applied to the *Hemiascomycetes* promoters delineated all GGTGGCAAA elements and degenerate nonamers thereof with maximally two base exchanges. These sequences (hits) were sorted by decreasing frequency and one (in a few cases, two) of these hits was selected for each gene that fulfilled the following criteria: (i) frequency ≥1; (ii) none, one, or two base exchanges, in this order; and (iii) the motif preferably conforming to the sequence DRTGGCRAN (i.e., leaving the "central" core of PACE unchanged). These criteria were built on the following three observations.

In our previous reports we observed that modification of the central GC or an exchange of the (central) C residue in PACE abolishes or reduces the binding of Rpn4p (Mannhaupt et al. 1999; Kapranov et al. 2001).

With four exceptions, the proteasomal genes from the *Saccharomyces* "sensu strictu" species (Table 2A, group 1) possess the unique sequence GGTGGCAAA (in direct or opposite orientation with the respective gene) in comparable distance upstream from the translational start site. This motif deviates by one nucleotide (GGTGGCGAA) in

Table 2 PACE and PACE-like motifs upstream of *Hemiascomycetes* genes

A Group 1 Species

Upstream of	*S. cerevisiae*		*S. paradoxus*		*S. mikatae*		*S. bayanus*		*S. kudriavzevii*	
RPT1	TTTGCCACC	154i	TTTGCCACC	154i	NA		NA		NA	
RPT2	GGTGGCAAA	130	GGTGGCAAA	132	NA		GGTGGCAAA	123	NA	
RPT3	GGTGGCAAA	175	GGTGGCAAA	178	NA		GGTGGCAAA	177	GGTGGCAAA	176
RPT4	TTTGCCACC	111i	TTTGCCACC	127i	NA		TTTGCCACC	112i	NA	
RPT5	GGTGGCAAA	163	GGTGGCAAA	155	NA		GGTGGCAAA	155	GGTGGCAAA	156
RPT6	GGTGGCAAA	83	GGTGGCAAA	82	GGTGGCAAA	84	GGTGGCAAA	101	GGTGGCAAA	78
RPN1	GGTGGCAAA	144	GGTGGCAAA	147	NA		GGTGGCAAA	144	NA	
RPN2	GGTGGCAAA	118	GGTGGCAAA	117	NA		GGTGGCAAA	117	NA	
RPN3	TTTGCCACC	94i	TTTGCCACC	94i	TTTGCCACC	94i	TTTGCCACC	92i	TTTGCCACC	94i
RPN5	GGTGGCAAA	137	GGTGGCAAA	139	NA		GGTGGCAAA	150	GGTGGCAAA	144
RPN6	GGTGGCAAA	112	too short		GGTGGCAAA	84	NA		NA	
RPN7	TTTGCCACC	153i	TTTGCCACC	154i	TTTGCCACC	153i	TTTGCCACC	167i	TTTGCCACC	161i
RPN8	**A**GTGGCAAA	163	**A**GTGGCAAA	163	NA		**A**GTGGCAAA	172	**A**GTGGCAAA	169
RPN9	GGTGGCAAA	108	GGTGGCAAA	112	GGTGGCAAA	103	GGTGGCAAA	145	GGTGGCAAA	104
RPN10	TT**C**GCCACC	104i	TT**C**GCCACC	104i	NA		TT**C**GCCACC	123i	TT**C**GCCACC	104i
RPN11	GGTGGCAAA	123	GGTGGCAAA	124	NA		GGTGGCAAA	148	NA	
RPN12	TTTGCCACC	96i	TTTGCCACC	93i	TTTGCCACC	93i	TTTGCCACC	105i	TTTGCCACC	102i
RPN13	GGTGGC**G**AA	117	GGTGGC**G**AA	120	GGTGGC**G**AA	120	GGTGGC**G**AA	122	GGTGGC**G**AA	123
PRE1	GGTGGCAAA	141	GGTGGCAAA	146	GGTGGC**G**AA	139	GGTGGCAAA	142	GGTGGCAAA	142
PRE2	GGTGGCAAA	154	GGTGGCAAA	155	NA		GGTGGCAAA	156	GGTGGCAAA	158
PRE3	GGTGGCAAA	117	GGTGGCAAA	298	GGTGGCAAA	299	GGTGGCAAA	307	GGTGGCAAA	299
PRE4	TTTGCCACC	110i	TTTGCCACC	111i	TTTGCCACC	108i	TTTGCCACC	104i	TTTGCCACC	105i
PRE5	**A**GTGGCAAA	147	**A**GTGGCAAA	148	**A**GTGGCAAA	149	**A**GTGGCAAA	154	**A**GTGGCAAA	159
PRE6	TTTGCCACC	108i	TTTGCCACC	109i	TTTGCCACC	108i	TTTGCCACC	109i	TTTGCCACC	106i
PRE7	GGTGGCAAA	86	GGTGGCAAA	87	GGTGGCAAA	86	GGTGGCAAA	88	GGTGGCAAA	87
PRE8	GGTGGCAAA	119	GGTGGCAAA	118	GGTGGCAAA	120	GGTGGCAAA	121	GGTGGCAAA	117
PRE9	GGTGGCAAA	158	GGTGGCAAA	158	GGTGGCAAA	160	GGTGGCAAA	159	NA	
PRE10	GGTGGCAAA	127	GGTGGCAAA	126	GGTGGCAAA	128	GGTGGCAAA	140	GGTGGCAAA	127
PUP2	GGTGGCAAA	103	GGTGGCAAA	104	GGTGGCAAA	104	GGTGGCAAA	103	GGTGGCAAA	86
PUP3	GGTGGCAAA	182	GGTGGCAAA	184	NA		GGTGGCAAA	212	GGTGGCAAA	186
UBA1	GGTGGCAAA	129	NA		NA		GGTGGCAAA	132	NA	
CDC48	GGTGGCAAA	141	GGTGGCAAA	143	GGTGGCAAA	144	GGTGGCAAA	143	GGTGGCAAA	143

B Group 2 Species

Upstream of	*S. kluyveri*		*K. waltii*		*S. castellii*		*A. gossypii*		*K. lactis*		*C. glabrata*	
RPT1	TT**C**GCCACC	227i	TTTGCCACC	302i	TTTGCCACC	107i	GGTGGCAAA	11	GGTGGC**G**AA	77	TTTGCCACC	352i
RPT2	GGTGGCAAA	345	GGTGGCAAA	71	GGTGGCAAA	108	GGTGGCAAA	92	**A**GTGGC**G**AA	158	**T**GTGGCAAA	420
RPT3	NA		GGTGGCAAA	116	GGTGGCAAA	140	ND		GGTGGCAAA	294	GGTGGCAA**G**	245
RPT4	NA		TT**C**GCCACC	86i	TTTGCCACC	106i	TTTGCCACC	102i	TTTGCCACC	215i	TTTGCCAC**A**	179i
RPT5	NA		GGTGGCAAA	42	GGTGGCAAA	88	GGTGGCAAA	13	GGTGGCAAA	201	TTTGCCAC**A**	140i
RPT6	GGTGGCAAA	75	GGTGGCAAA	66	GGTGGCAAA	90	GGTGGCAAA	66	**A**GTGGCAAA	198	**T**GTGGC**G**AA	158
RPN1	NA		GGTGGC**G**AA	103	GGTGGCAAA	121	GGTGGCAAA	114	GGTGGCAAA	207	**A**GTGGCAAA	298
RPN2	GGTGGCAAA TTTGCCACC	85 114i	GGTGGCAAA	58	GGTGGCAAA	190	GGTGGCAAA	83	GGTGGCAAA	126	GGTGGCAAA	411
RPN3	TTTGCCACC	62i	TTTGCCACC	11i	TTTGCCACC	98i	TTTGCCACC	79i	TTTGCCAC**T**	93i	TTTGCCACC	211i
RPN5	GGTGGCAAA	89	GGTGGCAAA	102	GGTGGCAAA	76	GGTGGCA**T**A	63	GGTGGCAAA	182	GGTGGC**G**AA	254
RPN6	GGTGGCAAA	71	GGTGGCAAA	53	NA		GGTGGCAAA	83	GGTGGCAAA	123	TTTGCCAC**T**	197i
RPN7	TTTGCCACC	65i	TTTGCCACC	52i	TTTGCCACC	47i	GGTGGCAAA	74	TTTGCCACC	82i	**A**GTGGCAAA	93
RPN8	GGTGGCAAA	54	GGTGGCAAA	71	GGTGGCAAA	165	GGTGGCAAA	99	**A**GTGGCAAA	251	GGTGGCAAA	243
RPN9	TTTGCCACC	59i	too short		GGTGGCAAA	82	GGTGGCAAA	101	GGTGGC**G**AA	85	TTTGCCAC**A**	245i
RPN10	TTTGCCACC	54i	TTTGCCACC	71	TTTGCCACC	48i	ND		**G**TTGCCACC	178i	GGTGGCAA**T**	329
RPN11	GGTGGCAAA	94	GGTGGCAAA	80	too short		GGTGGCAAA	105	**A**GTGGCAAA	264	**A**GTGGCAAA	285
RPN12	TTTGCCACC	62i	**A**GTGGCAAA	71	**A**GTGGCAAA	138	GGTGGCA**GC**	91	TTTGCCACC	42i	TTTGCCACC	256i
RPN13	GGTGGCAAA	64	GGTGGCAAA	89	GGTGGCAAA	70	GGTGGCAAA	49	ND		TTTGCCA**T**C	407i
PRE1	GGTGGCAAA	107	GGTGGCAAA	89	GGTGGCAAA	101	GGTGGCAAA	106	GGTGGCAAA	86	**A**GTGGC**G**AA	277
PRE2	GGTGGCAAA	46	GGTGGCAAA	81	GGTGGCAAA	105	GGTGGCAAA	87	**A**GTGGCAAA	146	GGTGGC**G**AA	158
PRE3	NA		GGTGGCAAA	214	GGTGGCAAA	158	GGTGGCAAA	88	GGTGGCAAA	124	**A**GTGGCAAA	217
PRE4	TTTGCCACC	78i	TTTGCCACC	67i	TTTGCCACC	80i	TTTGCCACC	39i	TTTGCCAC**T**	140i	TTTGCCAC**T**	344i
PRE5	GGTGGCAAA	77	GGTGGCAAA	104	**A**GTGGCAAA	116	AA**CA**CCACC	107i	GGTGGC**G**AA	151	**T**GTGGCAAA	150
PRE6	NA		TT**C**GCCACC	148i	TTTGCCACC	93i	TTTGCCACC	71i	TTTGCCACC	164i	GGTGGCAA**G**	217
PRE7	GGTGGCAAA	22	GGTGGC**G**AA	80	GGTGGCAAA	121	GGTGGCAAA	80	TTTGCCACC	91i	**A**GTGGCAAA	138
PRE8	GGTGGC**G**AA	375	GGTGGCAAA	393	**A**GTGGCAAA	84	GGTGGCAAA	74	GGTGGCAAA	124	**A**GTGGCAAA	138
PRE9	GGTGGCAAA	98	GGTGGC**G**AA	89	GGTGGCAAA	165	GGTGGCAAA		**A**GTGGCAAA	161	GGTGGCAA**C**	262
PRE10	GGTGGCAAA	396	GGTGGCAAA	136	GGTGGCAAA	137	GGTGGCAAA	98	ND		GGTGGCAA**G**	149
PUP2	NA		GGTGGCAAA	85	GGTGGCAAA	99	GGTGGC**G**AA	94	**A**GTGGCAAA	300	GGTGGC**GAC**	229
PUP3	TTTGCCACC	130i	**A**GTGGCAAA	222	GGTGGCAAA	137	GGTGGCAAA	273	**A**GTGGCAAA	112	GGTGGC**G**AA	433
UBA1	NA		GGTGGCAAA	68	GGTGGCAAA	118	GGTGGCAAA	108	**A**GTGGCAAA	134	**T**GTGGC**C**AA	462
CDC48	GGTGGCAAA	77	GGTGGCAAA	100	GGTGGCAAA	150	GGTGGCAAA	137	GGTGGCAAA	394	GGTGGCAAA	231

Table 2 continued

C Group 3 Species

Upstream of	*C. albicans*		*C. dubliniensis*		*D. hansenii*		*Y. lipolytica*	
RPT1	**A**TTGCCAC**T**	84i	**A**TTGCCAC**T**	81i	G**AG**GGCAAA	85	**A**TTGCCACC	89i
							G**T**TGCCACC	128i
RPT2	GGTGGC**G**A**G**	59	GGTGGC**G**A**G**	68	G**T**TG**A**CAAA	477	GGTGGCAAA	209
RPT3	G**AA**GGCAAA	104	G**AG**GGCAAA	106	**A**TTGCCAC**T**	165i	TTT**C**CCACC	132i
RPT4	TTTGCCAC**T**	109i	TTTGCCAC**T**	112i	TTTGCCAC**T**	137i	C**T**CGCCACC	122i
RPT5	GGTGGCAA**C**	89	GGTGGCAA**C**	101	**A**GTGGCAA**G**	171	**A**GTGGCAA**T**	173
RPT6	**A**GTGGCAAA	81	GGTGG**T**AAA	20	G**AA**GGCAAA	56	**A**T**C**GCCACC	302i
RPN1	TTTGCCAC**T**	238i	TTTGCCAC**T**	192i	TTTGCCA**T**C	183i	**A**TTGCCAC**T**	41i
RPN2	GGTGGCAA**C**	199	GGTGGCAA**C**	214	GGTGGCAA**T**	94	GGTGGCAAA	82
RPN3	G**T**CGCCACC	115i	G**T**TGCCACC	133i	GGTGGCAA**C**	88	**A**GTGGC**G**AA	390
RPN5	G**T**TGGC**C**AA	155	TT**C**GCCACC	29i	C**T**TGCCACC	74i	T**G**TG**G**CACC	70i
RPN6	ND		ND		G**A**TGGCAAA	65	**T**GTGGCA**C**A	419
RPN7	TTTGCC**TT**C	55i	TTTGCC**TT**C	55i	TTTG**A**CA**T**C	406i	G**T**TGCCACC	140i
RPN8	G**AA**GGCAAA	138	G**AA**GGCAAA	135	TTTGCCAC**A**	9i	GGTGGCAA**C**	93
RPN9	G**A**TGGCAA**G**	96	TTT**T**CCACC	152i	GGTGGCAA**G**	52	**A**TTGCCACC	49i
RPN10	**A**GTGGCAA**T**	107	**A**GTGGCAA**T**	96	GGTGGC**G**A**T**	177	GGTGGCAAA	129
RPN11	TTTGCCAC**A**	14i	TTTGCCAC**A**	15i	C**T**TGCCACC	168i	TTTGCCAC**A**	64i
RPN12	GGTGGCAA**T**	94	GGTGGCAA**T**	87	GGTGGCAA**T**	84	GGTGGCAA**T**	205
RPN13	GGTG**T**CAAA	284	GGTGG**G**AA**T**	385	GGTGG**AA**A**T**	221	TTT**T**CCACC	72i
PRE1	G**AA**GGCAAA	58	G**AA**GGCAAA	61	G**A**TGGCAAA	205	C**T**TGCCACC	73i
PRE2	GGTGGCAA**C**	162	GGTGGCAA**C**	146	GGTGGCAA**G**	185	**T**GTGGCAA**C**	429i
PRE3	**A**GTGGCAAA	59	**A**GTGGC**G**AA	56	GGTGGCAA**T**	68	ND	
PRE4	**A**TTGCCAC**T**	136i	**A**TTGCCAC**T**	112i	TTTGCCAC**T**	149i	GGTGGC**G**AA	287
	G**AA**GGCAA**T**	143	G**AA**GGCAA**T**	119				
PRE5	**A**GTGGCAAA	82	**A**GTGGCAAA	88	ND		TT**C**GCCAC**A**	126i
PRE6	G**AA**GGCAAA	372	NA		TTTG**T**CACC	261i	TT**C**GCCACC	56i
PRE7	ND		**A**GTGGC**G**AA	187	ND		**A**T**C**GCCACC	36i
							AGTGGCAA**T**	25
PRE8	G**AA**GGCAAA	315	G**AA**GGCAAA	326	G**AA**GGCAAA	271	TT**C**GCC**T**CC	85i
PRE9	G**T**TGCCACC	217i	G**T**TGCCACC	184i	C**T**TGCCACC	80i	GGTGGC**G**AA	470
PRE10	**A**GTGGCAA**T**	300	**A**GTGGCAA**T**	296	GGTGGC**G**A**T**	312	**A**TTGCCACC	56i
PUP2	G**T**TGCCACC	114i	GGTGG**TC**AA	154	CTTGCCAC**T**	112i	**A**TTGCCACC	28i
PUP3	TTTGCCA**T**C	125i	NA		**AA**TGGCAAA	185	TTTGCCACC	122i
UBA1	TTTGCCAC**A**	32i	TTTGCC**TT**C	181i	**A**GTGGCAA**T**	59	C**T**TGCCAC**A**	44i
CDC48	GGTGGCAAA	397	GGTGGCAAA	389	GGTGGC**G**AA	149	**T**GTGGCAAA	44

Note. Nucleotide exchanges toward the "genuine" PACE box (GGTGGCAAA) are indicated in boldface. Numbers refer to positions of the boxes upstream from the translational start site; i indicates the occurrence of a box on the opposite strand. NA, promoter sequence not available; ND, no element detected. Elements conforming to the sequence GGTGGCAAA or GGTGGCAAN are highlighted by shading

the promoters of *RPN10* and *RPN13*, respectively, and by the first nucleotide (AGTGGCAAA) in the promoters of *PRE5* and *RPN8*, respectively. However, in the light of earlier microarray expression profiles (Eisen et al. 1998), these alterations seem to be tolerated, i.e., these modified boxes should act as functional *cis*-regulatory elements in binding Rpn4p.

Furthermore, Jelinsky and colleagues (2000) have delineated groups of coregulated genes in *S. cerevisiae* whose upstream regions bear specific regulatory sequence motifs. They observed that one group of coregulated genes contained a number of DNA excision repair genes and a large selection of protein degradation genes. Moreover, transcription of these genes was found to be modulated by Rpn4p, most likely via its binding to *MAG1* upstream repressor sequence elements (GGTGGCGA), which turned out to be almost identical to the proteasome-associated control element (PACE). The authors' statement "that the *MAG1* element normally behaves as a repressor binding site does not necessarily exclude Rpn4p's behaving as an activator at this site" may be taken as a further indication that GGTGGCGA can act as a functional *cis*-regulatory element in binding Rpn4p.

Our results are outlined schematically in Fig. 2 and detailed in Table 2. As can be inferred from Table 2A, *S. paradoxus*, the species most closely related to *S. cerevisiae*, exhibits an identical "PACE pattern." Remarkably, also the upstream positions of the motifs are very similar, if not identical, to each other. The "PACE pattern" changes only minimally within the *Saccharomyces* "sensu stricto" group, as far as we can conclude from the sequences available in the databases.

In group 2 (Fig. 2, Table 2B), we observed the occurrence of a genuine PACE element for the majority of the proteasomal genes, though the upstream positions of these elements are much more variable compared to those in group 1. Further, there is an increasing number of cases in group 2 (particularly for *K. lactis* and *C. glabrata*), in which the PACE element is presumably substituted by either AGTGGCAAA (change of G to an A residue in position 1) or GGTGGCGAA, or even, in one case, by a PACE-like sequence with two base exchanges conforming

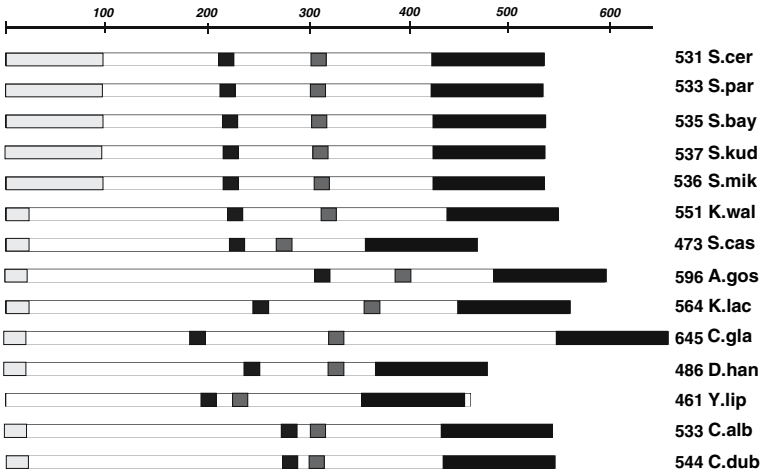

Fig. 1 Schematic representation of conserved domains in the Rpn4p homologues from *Hemiascomycetes* species. Lengths of proteins (in amino acid residues) are indicated. Black boxes, atypical Zn-finger; gray boxes, acidic domains 1 and 2; light-gray boxes, N-terminally conserved sequences. For more details, see Supplement 1

to the above rules. In *C. glabrata*, we observe the "canonical" PACE element in only 5 cases among the 32 promoter sequences. But interestingly, *C. glabrata* exhibits an element (TGTGCCAAA) similar to AGTGGCAAA six times. The "alternative" PACE element GGTGGCGAA is present in *C. glabrata* three times.

The "PACE patterns" of the group 3 species (Fig. 2, Table 2C) exhibit still greater variations than those of group 2. In *Y. lipolytica* the "canonical" PACE element is found in four cases, while *C. albicans* and *C. dubliniensis* exhibit this sequence only for *CDC48* (see below); none occurs in *D. hansenii.* However, instead we observed again a number of elements in which position 1 has been changed to an A residue: AGTGGCAAA occurs four times in *C. albicans* and three times in *C. dubliniensis* and *D. hansenii*, respectively; no such element is present in *Y. lipolytica*. The "alternative" PACE element GGTGGCGAA occurs only once in *D. hansenii* and *Y. lipolytica*, respectively, and not in the other members of group 3.

In addition, one observation we paid particular attention to is that in *K. lactis*, *C. glabrata*, and the group 3 PACE patterns, increasing numbers of cases are found in which PACE-like elements with one base exchange at their 3′-ends (GGTGGCAAN) occur (see Tables 2B and C). The figures are as follows: 1 in 32 for *K. lactis*; 5 in 32 for *C. glabrata*; 7 in 32 for *C. albicans* and *C. dubliniensis*, respectively; 8 in 32 for *D. hansenii*; and 9 in 32 for *Y. lipolytica*. Among these species, we also observed a number of PACE-like elements with variations in both position 1 and position 9; i.e., only the seven core positions have been conserved. For example, the occurrence of AGTGGCAAN is 4 in 32 for *C. albicans and C. dubliniensis*, respectively; 3 in 32 for *D. hansenii*; and 4 in 32 for *Y. lipolytica*. A PACE-like element conforming to the sequence GAAGGCAAA (i.e., changes in positions 2 and 3 vs. the canonical element as reported by Gasch et al. [2004]) is present at 5 in 32 in *C. albicans* and *C. dubliniensis*, respectively, and 2 in 32 in *D. hansenii*, but zero times in *Y. lipolytica*. These findings are discussed below.

Discussion

We performed a search in 15 *Hemiascomycetes* species (see Methods) to exploit the evolutionary maintenance of the transcription factor Rpn4p together with the so-called PACE element, which initially has been identified as an Rpn4p binding site for the majority of the proteasomal and a number of additional genes in *S. cerevisiae* (Mannhaupt et al. 1999; Kapranov et al. 2001). Thus, in extension to the 10 species analyzed by Gasch et al. (2004), we were able to add 2 species (*C. glabrata* and *K. lactis*) with an intermediate phylogenetic relationship to the *Saccharomyces* and 3 species (*C. dubliniensis*, *D. hansenii*, and *Y. lipolytica*) with a more distant phylogenetic relationship to the *Saccharomyces* species (Dujon 2006). As an "outgroup" in our searches, we have used the corresponding sequences from *S. pombe*, *N. crassa*, two *Aspergillus* species, *Arabidopsis*, and human (see Methods).

In accordance with earlier notions (e.g., Wolf and Hilt 2004) we found that the 30 proteasomal gene products considered here as well as Uba1p and Cdc48p are highly conserved throughout all these species (Table 1), as they are of fundamental importance for cellular function in eukaryotes. Further, the domain structures of the Rpn4p homologues in the *Hemiascomycetes* are well conserved. The highest similarity is observed for the C-terminal portions (ca. 130 residues) in which the domain of the atypical Zn-finger is located (Fig. 1). CLUSTAL W resulted in a nearly perfect alignment of the sequences in this region (Supplement 1). By contrast, the acidic domains reveal a greater divergence, except those among the *S. cerevisiae* "sensu stricto" species. However, by CLUSTAL W

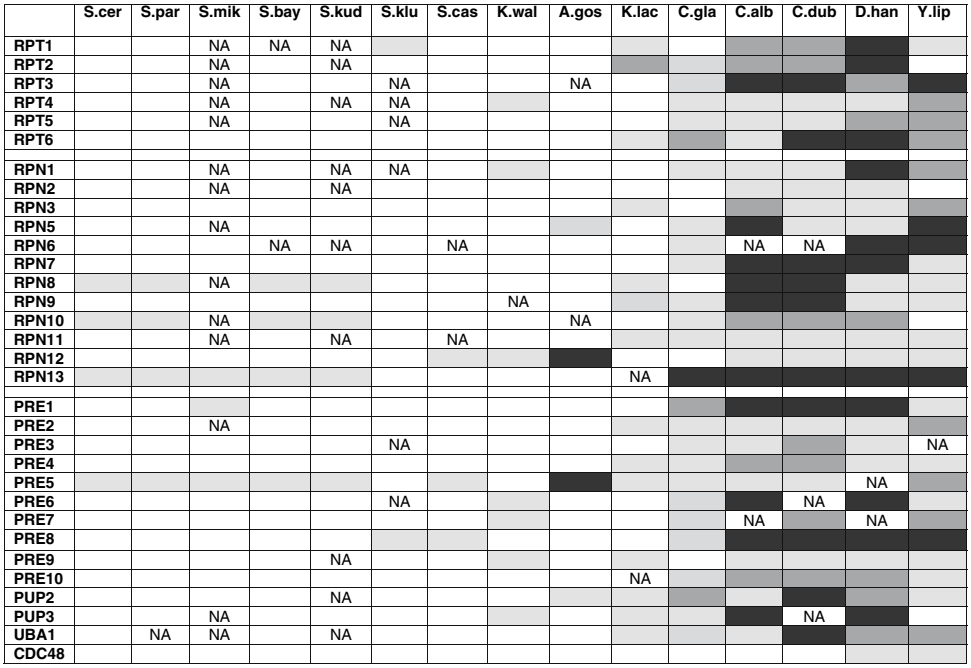

Fig. 2 Schematic representation of the occurrence of PACE and PACE-like motifs in the upstream regions of proteasomal genes in *Hemiascomycetes* species. White box, genuine PACE, GGTGGCAAA; light-gray box, motif with one base exchange vs. PACE, conforming to DGTGGCRAN; gray box, motif with two base exchanges vs. PACE, conforming to DGTGGCRAN; dark-gray box, two base exchanges not conforming to DGTGGCRAN. NA, upstream sequence not available. For more details, see Table 2

alignments and by eye inspection, two stretches rich in acidic amino acids, as well as similar in length and relative distances, are present in the residual Rpn4p-like sequences. As the acidic domains will probably function as activating domains, they need not be as strictly conserved as DNA-binding domains like Zn-fingers. Thus the high similarity of the acidic modules in Rpn4p of the "sensu stricto" species reflects their close evolutionary relationship, while the greater variation of these modules among the other species is likely to be a consequence of much greater evolutionary distances.

A microarray-based genomic survey had revealed that the *S.cerevisiae* proteasomal gene cluster exhibits a "stereotypical" expression pattern under varying environmental conditions (Eisen et al. 1998). Moses and colleagues have convincingly shown that functional PACE elements are evolutionary maintained in the upstream regions of those proteasomal genes from the *Saccharomyces* "sensu strict" group that follow the "stereotypical" expression pattern (Moses et al. 2003, 2004).

In our hands, conventional benchmarking tools (Pollard et al. 2004) such as DiAlign or CLUSTAL W (data not shown) allowed for the detection of PACE or PACE-like sequences in the upstream promoters of most proteasomal genes from the "sensu stricto" species, while extending such searches to the group 2 and group 3 species was hampered by the fact that during evolution a greater extent of rearrangements (including deletions/insertions) among homologous genes and their flanking sequences in general has occurred (e.g., Dujon et al. 2004; Fischer et al. 2006). When we tested DiAlign or CLUSTAL W to align the 500-bp upstream sequences from the 15 *Hemiascomycetes*, even those sequences that enharbor a unique PACE element were not correctly aligned. These routines also largely failed in pairwise alignments. However, when we preselected 50 bp including the presumptive elements detected by our JAVA program, these were correctly aligned by CLUSTAL W, at the same time demonstrating that the sequences flanking the elements share little or no similarity except those of the "sensu stricto" (group 1) species.

While Gasch et al. (2004) have chosen statistical approaches to build "meta-matrices" for PACE-like upstream elements, the simple routine we developed basically lead to similar results. The criteria we applied to the selection of the motifs in our approach (see Results) were based on earlier findings and are in agreement with an important hypothesis of the comparative genomics paradigm stating that as evolutionary distance increases, observing a match with a given level of conservation should become less and less likely by chance. Moses and collaborators (2003, 2004) have characterized the evolution of experimentally validated transcription factor binding sites (TFBSs) in the *Saccharomyces cerevisiae* genome, finding that functional TFBSs evolve more slowly than flanking intergenic regions, pointing to a purifying selection of such elements. They concluded that as evolutionary distance increases, one would expect fewer matches to a given matrix to be conserved by chance. Although not every functional binding site will remain under purifying selection, as a result of either functional change or binding-site turnover, a large subset of functional binding sites does

remain under purifying selection. These authors also pointed out that there might be considerable position-specific variation in evolutionary rates within TFBSs. They further showed that evolutionary rate at each position is a function of the selectivity of the factor for bases at that position. We paid attention to this finding in that we considered rather exclusively PACE-like motifs which have kept the core of PACE (see Results), which seems to be essential for function.

In our approach, we explicitly listed the putative elements and their upstream locations for 30 proteasomal genes and 2 genes (*UBA1* and *CDC48*) which are under the control of Rpn4p (including the five recently sequenced *Hemiascomycete* species). Comparisons thus allowed for a more detailed assessment of the relationships between the respective elements. Evaluating Table 2 immediately implies that during evolution of the *Hemiascomycetes* there is a continuous enrichment in the number of genuine PACE elements, so that the earlier and more "degenerate" PACE-like motifs but functional with their cognate factors could have converged to "canonical" PACE motifs in the evolutionary most recent species. In a large number of incidences, convergence could have been brought about by a single base change in preexisting motifs in the ancestors, notably in position 1, 9, or 7 of the sequence (mainly conforming to AGTGGCAAA, GGTGGCAAN, or GGTGGCGAA, respectively). This is obvious, for example, in comparing the motifs found in *K. lactis* to the other species in group 2 and those in group 1. A similar notion is valid for the motifs found in *C. glabrata*, because there are more "degenerate" PACE elements in *C. glabrata* than in the rest of species in group 2 or 1. A peculiarity of *C. glabrata* is the repeated occurrence of TGTGGCAAA (see Results), whereby a single base exchange would result in the "canonical" motif. Thus, it appears that the patterns in group 2 do not exactly reflect the divergence times as worked out for the phylogenetic tree on other criteria (e.g., Fisher et al. 2006; Dujon, 2006; Scannel et al. 2006).

Compared to group 2, there are a larger number of motifs in group 3 that conform to the sequences AGTGGCAAN or GGTGGCAAN. Examples that indeed mutations occur in these motifs in closely related species can be seen for *C. albicans* and *C. dubliniensis*. The upstream region of *PRE3* contains the motif AGTGGCAAA in *C. albicans*, whereas it reads AGTGGCGAA in *C. dubliniensis*. The upstream region of *RPN3* exhibits the motif GGTGGCAAC in *C. dubliniensis*, and GGTGG CGAC in *C. albicans*, in nearly identical upstream locations.

In any case, provided that all of the proteasomal genes in the various species are subject to regulation by their cognate Rpn4 factors, these would have to be flexible enough to bind degenerate PACE-like sequences in these species. Implicitly, this has been demonstrated for two "extreme" species, *S. cerevisiae* and *C. albicans*, by in vitro binding and competition experiments (Gasch et al. 2004), for which oligonucleotides comprising the decamers GGTGG CAAAA (Sequence A), AGTGGCAACA (Sequence C), and GAAGGCAAAA (Sequence B) were used. While *C. albicans* Rpn4 bound to these with comparable efficiency, *S. cerevisiae* Rpn4 preferentially bound to GGTGG CAAAA and less to AGTGGCAACA but had a reduced ability to bind to GAAGGCAAAA. These authors also stated that *S. cerevisiae* Rpn4p could transcribe a reporter gene to higher levels if Sequence A was present in its promoter compared to when Sequence B or a minimal promoter was placed upstream of the reporter gene, pointing out that *S. cerevisiae* Rpn4p (and probably their closest relatives) largely lost the ability to bind productively to Sequence B.

The authors have called Sequence B "the *C. albicans* specific element," but it may be noted that analogous motifs occur also in *C. dubliniensis* (at similar upstream locations in the same proteasomal genes as in *C. albicans*) but also in *D. hansenii*. We find that for the two *Candida* species in group 3, the occurrence is $\sim$17 % among the proteasomal upstream elements and only $\sim$7 % for *D. hansenii*, while we wish to emphasize that the majority of the elements throughout the group 3 species still fit into the matrices GGTGGCAAN or AGTGGCAAN (base alterations in position 9).

Inspection of all proteasomal gene upstream regions in *C. albicans* reveals that only 8 of 30 of the elements fully conform to the above decamers, while the rest exhibit one or two base exchanges, again supporting the notion that Rpn4p must be flexible enough to bind degenerate PACE-like sequences, if the complete set of *cis*-regulatory elements is used in the regulatory network. It may well be that nonamers with a conserved core and one or two base variations might suffice for Rpn4 binding. Note that nearly all of our assignments and also the matrices formulated by Gasch et al. (2004) point to the significance of the central –GGC–. Gasch et al. (2004) argued that the different binding specificities found between *S. cerevisiae* and *C. albicans* reside in the second Zn-finger of the Rpn4 homologues which is proposed to contact the first half of the DNA-binding site. Given the possible significance of the core part of the PACE-like elements, we may speculate that this part is contacted by the first (atypical) finger, which in our alignment is found to be equal in length, very highly conserved in its "loop" region (12 residues) between $CX_{10}C$ and HX_4H, and highly conserved in the linker region to the second Zn- finger, in all *Hemiascomycetes*. Unfortunately, it is unknown which parts of the atypical Zn-finger may make contact to which nucleotides of the *cis*-regulatory elements, and this remains largely

unpredictable by comparisons among the Rpn4 homologues, as no models have been developed for such an atypical Zn-fingers other than for conventional Zn-fingers (S. Wolfe et al. 2000; Pabo et al. 2001).

At first view, the finding of Gasch et al. (2004) that the *N. crassa* Rpn4 homologue can bind the above decamers comes as a surprise. A comparison between the respective DNA-binding domains reveals that there is again high conservation within the first $CX_{10}C/HX_4H$ domain and the second $CXXC/HX_4H$ domain, while the region between these two domains is shorter by seven amino acids in *N. crassa* compared to the *Saccharomyces* species. (This observation is also valid for the Rpn4p homologues from *S. pombe*, *A. nidulans*, and *A. fumigatus*; cf. Supplement 2.) Thus the DNA-binding domain in *N. crassa* may still allow for binding of PACE sequences. However, that none of these Rpn4 homologues contain acidic domains and no PACE-like elements are found upstream of the proteasomal genes argues for the proposition that no regulatory networks as in the *Hemiascomycetes* do exist in the other fungi.

Earlier and more degenerate PACE-like motifs, but functional with their cognate factors, have converged to "canonical" PACE motifs in the species that in evolution have separated more recently, such as the group 1 species, in which practically no changes have occurred either in the PACE patterns or in the DNA-binding sites of Rpn4p. Obviously, the ability of Rpn4p of the more recently segregated species to bind to degenerate PACE-like motifs has been reduced by adapting the Rpn4p sequences concomitantly (Gasch et al. 2004). This scenario would probably afford a stepwise (mutational) convergence of "pre-PACE" elements into "true" PACE motifs as evolution proceeded and would be in agreement with the proposal that *cis*-regulatory changes are an important source of genetic variation (Wray et al. 2003) and that gains (and losses) of functional binding sites significantly contribute to these changes (e.g., Dermitzakis and Clark, 2002). Likewise, as can be inferred from the comparison of the Rpn4 proteins (Supplement 1), mutations in the highly conserved DNA-binding domains must have contributed to their capability of interacting with the actually occurring PACE-like *cis*-regulatory elements. We have paid attention to this by putting the Rpn4p homologue of group 1 and 2 into an order that parallels the variability in the cognate elements found in these species: the most pronounced alterations in the Rpn4p DNA-binding domains are observed for those species in which the PACE-patterns exhibit the greatest variety (group 2), while the restriction of binding to a more specialized PACE element such as GGTGGCAAA in the more recently segregated *Saccharmoyces* (group 1) is mirrored by considerably fewer or no alterations in the DNA-binding domains. For the group 3 species, it is difficult to establish such a strict correlation.

Overall, it seems evident that the concerted regulation of the proteasomal genes by Rpn4 proteins is a special acquisition of the *Hemiascomycetes*, but that no similar mechanism is operative in *S. pombe*, other fungi, or higher eukaryotes. This notion is substantiated by investigations on the ubiquitin-proteasomal network from *Drosophila* (e.g., Wojcik and DeMartino 2002; Lundgren et al. 2005) or mammalian cells (e.g., Meiners et al. 2003) that have clearly demonstrated the existence of a concerted regulation but have not identified a system similar to the one in yeast. We hope that our observations will stimulate further experiments to better understand the regulatory network of this most important system for cell viability, in *Hemiascomycetes* as well as in higher eukaryotes.

Acknowledgment The Java Programme for Pattern Search was kindly developed by H. R. Mannhaupt.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Aspergillus Sequencing Project (2003) Aspergillus Sequencing Project. Broad Institute, MIT and Harvard, Cambridge, MA

Chiang DY, Moses AM, Kellis M, Lander ES, Eisen MB (2003) Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. Genome Biol 4:R43

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 301:71–76

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19:1114–1121

Dietrich FS, Voegeli S, Brachat S, Choi S, et al. (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304:304–307

Dujon B (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet. 22:375–387

Dujon B, Sherman D, Fischer G, et al. (2004) Genome evolution in yeasts. Nature 430:35–44

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868

Fischer G, Neuveglise C, Durrens P, Gaillardin C, Dujon B (2001) Evolution of gene order in the genomes of two related yeast species. Genome Res 11:2009–2019

Fischer G, Rocha EP, Brunet F, Vergassola M, Dujon B (2006) Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genet 2:e32

Galagan JE, Calvo SE, Borkovitch KA, et al. (2003) The genome sequence of the filamentous fungus Neurospora crassa. Nature 422:859–868

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4241–4257

Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biol 2:e398

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. (1996) Life with 6000 genes. Science 274:563–567

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296:1205–1214

Jelinsky SA, Estep P, Church GM, Samson LD (2000) Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. Mol Cell Biol 20:8157–8167

Jones T, Federspiel NA, Chibana H, et al. (2004) The diploid genome sequence of Candida albicans. Proc Natl Acad Sci USA 101:7329–7334

Ju D, Xie Y (2004) Proteasomal degradation of RPN4 via two distinct mechanisms, ubiquitin-dependent and -independent. J Biol Chem 279:23851–23854

Kapranov AB, Kuriatova MV, Preobrazhenskaia OV, Tiutiaeva VV, Stuka R, Feldmann H, Karpov VL, Karpov V (2001) Isolation and identification of PACE-binding protein rpn4—a new transcription activator, participating in regulation of 26S proteosome and other genes. Mol Biol (Mosk) 35:420–431

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617–624

Llorente B, Malpertuy A, Neuvéglise C, et al. (2000) Genome rearrangements in yeasts. Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with Saccharomyces cerevisiae. FEBS Lett 487:101–112

Lundgren J, Masson P, Mirzaei Z, Young P (2005) Identification and characterization of a Drosophila proteasome regulatory network. Mol Cell Biol 25:4662–4675

Makalowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA 95:9407–9412

Malpertuy A, Tekaia F, Casaregola S, et al. (2000) Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. FEBS Lett 487:113–121

Mannhaupt G, Schnall R, Karpov V, Vetter I, Feldmann H (1999) Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast. FEBS Lett 450:27–34

Meiners S, Heyken D, Weller A, Ludwig A, Stangl K, Kloetzel K-P, Kruger E (2003) Inhibition of proteasome activity induces concerted expression of proteasome genes and de novo formation of mammalian proteasomes. J Biol Chem 278:21517–21525

Morgenstern B, Prohaska SJ, Poehler D, Stadler PF (2006) Multiple sequence alignment with user-defined anchor points. Algorithms Mol Biol 1:6

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB (2003) Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol Biol 3:19

Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5:R98

Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, et al. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature 438:1092–1093

Owsianik G, Balzi L, Ghislain M (2002) Control of 26S proteasome expression by transcription factors regulating multidrug resistance in Saccharomyces cerevisiae. Mol Microbiol 43:1295–1308

Pabo CO, Peisach E, Grant RA (2001) Design and selection of novel Cys2His2 zinc finger proteins. Annu Rev Biochem 70:313–340

Pollard DA, Bergman CM, Stoye J, Celniker S, Eisen MB (2004) Benchmarking tools for the alignment of functional noncoding DNA. BMC Bioinform 5:6

Prinz S, Avila-Campillo I, Aldridge C, Srinivasan A, Dimitrov K, Siegel AF, Galitski T (2004) Integrated molecular network control. Genome Res 14:380–390

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341–345

Souciet JL, Aigle M, Artiguenave F, et al. (2000) Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett 487:3–12

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res 31:3593–3596

Wang L, Mao X, Ju D, Xie Y (2004) Rpn4 is a physiological substrate of the Ubr2 ubiquitin ligase. J Biol Chem 279:55218–55223

Wolf DH, Hilt W (2004) The proteasome: a proteolytic nanomachine of cell regulation and waste disposal. Biochim Biophys Acta 1695:19–31

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Wolfe SA, Nekludova L, Pabo CO (2000) DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 29:183–212

Wojcik C, DeMartino GN (2002) Analysis of Drosophila 26 S proteasome using RNA interference. J Biol Chem 277:6188–6197

Wood V, Gwilliam R, Rajandream MA, et al. (2002) The genome sequence of Schizosaccharomyces pombe. Nature 415:871–880

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20:1377–1419

Xie Y, Varshavsky A (2001) RPN4 is a ligand, substrate, and transcriptional regulator of the 26S proteasome: a negative feedback circuit. Proc Natl Acad Sci USA 98:3056–3061