PROTEINS: Structure, Function, and Genetics 13:41–56 (1992)

Comparative Analysis of the β Transducin Family With Identification of Several New Members Including *PWP1*, a Nonessential Gene of *Saccharomyces cerevisiae* That Is Divergently Transcribed From *NMT1*

Robert J. Duronio,[1] Jeffrey I. Gordon,[1] and Mark S. Boguski[2]
[1]Department of Molecular Biology and Pharmacology, Washington University School of Medicine, St. Louis, Missouri 63110, and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

ABSTRACT **While investigating the expression of the *Saccharomyces cerevisiae* myristoyl-CoA:protein N-myristoyltransferase gene (NMT: E.C. 2.3.1.97) by Northern blot analysis, we observed another RNA transcript whose expression resembled that of *NMT1* during meiosis and was derived from a gene located <1 kb immediately upstream of *NMT1*. This new gene, designated *PWP1* (for periodic tryptophan protein), is divergently transcribed from *NMT1* and encodes a 576-residue protein. Null mutants of *PWP1* are viable, but their growth is severely retarded and steady-state levels of several cellular proteins (including at least two proteins that label with exogenous [^{3}H]myristic acid) are drastically reduced. New methods for database searching and assessing the statistical significance of sequence similarities identify *PWP1* as a member of the β-transducin protein superfamily. Two other previously unrecognized β-transducin-like proteins (*S. cerevisiae MAK11* and *D. discoideum* AAC3) were also identified, and an unexpectedly high degree of sequence homology was found between a *Chlamydomonas* β-like polypeptide and the C12.3 gene of chickens. A systematic and quantitative comparative analysis resulted in classifying all β-transducin-like sequences into 11 nonorthologous families. Based on specific sequence attributes, however, not all β-transducin-like sequences are expected to be functionally similar, and quantitative criteria for inferring functional analogies are discussed. Possible roles of repetitive tryptophan residues in proteins are also considered.**

Key words: fungal genes, molecular cloning, molecular sequence data, gene expression, database searching, sequence alignment, repetitive sequences, signal transduction, protein N-myristoylation

INTRODUCTION

β-transducin (also known as G_β) is one of three subunits of the heterotrimeric ($G_\alpha G_\beta G_\gamma$) GTP-binding proteins that act as intermediaries in transmembrane signal transduction in a variety of physiological contexts, including vision and hormone action.[1] The α subunit (G_α) belongs to a superfamily of GTPases that also includes *ras* oncogenes and proto-oncogenes ($p21^{ras}$) and translation elongation factors (e.g., EF-Tu) used in protein synthesis.[2,3] The biochemical functions of the β and γ subunits are less clear, but they are both necessary for membrane anchoring and receptor recognition and may have a regulatory role in G_α deactivation.[1] Shortly after the first β-transducin sequence (from bovine retina) became available, database searching revealed marked sequence similarity with the product of a cell division cycle gene (*CDC4*) of yeast,[4] the biochemical function of which remains unknown.[5] Since that time, a number of other β-transducin homologs have been identified,[6–12] but these sequences have shed little light on either the biochemistry of β-transducin itself or on the functions of proteins to which it is evolutionarily related.

Myristoyl-CoA:protein N-myristoyltransferase (NMT) catalyzes the co-translational attachment of myristate via an amide bond to the NH_2-terminal Gly_2 residue of a number of eukaryotic and viral proteins.[13,14] These include proteins involved in regulation of cell growth and differentiation (e.g., $p60^{src}$, G_α, the catalytic subunit of protein kinase-A), and the gag polyprotein precursors of several mammalian retroviruses (e.g., the $Pr55^{gag}$ of HIV-I) and the capsid proteins of a variety of picornaviruses (e.g., VP4 of poliovirus) and papovaviruses

Received March 31, 1991; revision accepted July 15, 1991.
Address reprint requests to Dr. Mark S. Boguski, National Center for Biotechnology Information, National Library of Medicine, NIH, Bldg. 38A, Room 8S-810, 8600 Rockville Pike, Bethesda, MD 20894.

(e.g., VP2 of polyoma virus and simian virus 40). The unique specificity of NMT for myristate[13,15,16] suggests that the physicochemical properties of this rare fatty acid (<2% of total cellular fatty acid in *S. cerevisiae*)[17] are needed for full expression of the biological function of some or all N-myristoylproteins. *NMT1* has been meiotically mapped by standard tetrad analysis to chromosome XII, 22 centimorgans (cM) centromere proximal to the CDC42 locus.[18] Insertional mutagenesis of this single copy gene causes recessive lethality, indicating that N-myristoylation of one or more *S. cerevisiae* proteins is necessary for vegetative growth.[19] The NH_2-terminal sequence motifs of several cloned yeast genes suggest that they should be substrates for NMT. These include the G_α subunit homolog *GPA1* (or *SCG1*),[20,21] which is involved in mating factor signal transduction[22]; the *ARF1* gene product,[23] which may be involved in intracellular transport within the Golgi complex[24]; and VPS15, a serine/threonine kinase involved in vesicle transport.[25]

During an analysis of the expression of the *S. cerevisiae NMT1* locus, we identified a new gene, *PWP1*, that is divergently transcribed from *NMT1* and encodes a nonessential protein with statistically significant sequence similarities to the G_β family of proteins. A comprehensive study of β-transducin homologs was performed, and the resulting classification provided a benchmark for making functional inferences. Consideration of sequence attributes suggests that the *PWP1* gene product may have nucleic acid-binding properties. The observed phenotype of *PWP1* null alleles suggests a regulatory role for this gene in cell growth or transcription.

MATERIALS AND METHODS

NMT1 and *PWP1* Gene Expression

Blots of total RNA derived from mitotically and meiotically synchronous cell cultures were prepared as previously described.[26] These blots were a kind gift of Glenn Bauer and Peter Burgers (Department of Biochemistry and Molecular Biophysics, Washington University). They were probed with ^{32}P-labeled DNA fragments encompassing different regions of the *NMT1* and *PWP1* genes (see under Results). Hybridization occurred at 65°C in a solution consisting of 1M NaCl, 1% sodium dodecyl sulfate (SDS), and 10% dextran sulfate. The final washing stringency was 65°C in Tris–HCl pH 7.5 prior to reprobing.

Cell Labeling and Quantitation of NMT Protein and Specific Activity

YM2061 and YB188 cells were grown to an OD_{600} of 1.5 in YPD broth at 30°C. Aliquots of the culture were metabolically labeled with either [9,10^{3}H]myristic acid (39.3 Ci/mmol, New England Nuclear, 100 μCi/ml of culture) or [^{35}S]methionine (1,149 Ci/mmol, New England Nuclear, 5 μCi/ml of culture) for 60 min at 30°C. After washing the cells once with phosphate-buffered saline (PBS) (1.37 M NaCl, 44 mM Na_2HPO_4, 17 mM KH_2PO_4, pH 7.3), lysates were prepared by boiling cells for 10 min in 240 mM Tris–HCl pH 6.8, 2% SDS, 0.4% β-mercaptoethanol, and 10% glycerol. Cellular debris were removed by subsequent centrifugation at 12,000*g* for 10 min. An equal amount of lysate protein from each time point (100 μg) was subjected to electrophoresis through a 12% polyacrylamide gel containing SDS (0.1%), which was then treated with En^{3}HANCE (New England Nuclear). Crude lysates were also prepared for use in Western blots and the determination of NMT specific activity. To do this, cells were vortexed in the presence of 500-μm glass beads in a solution containing 1X RIPA buffer (50 mM Tris–HCl pH 7.4, 150 mM NaCl, 20 mM EDTA, 0.1% SDS, 1% Triton X-100, 1% sodium deoxycholate, 10 μg/ml leupeptin, 10 μg/ml soybean trypsin inhibitor, 1 mM phenylmethylsulfonyl fluoride). Three 1-min periods of vortexing were interspersed with 2-min periods on ice, followed by centrifugation at 12,000*g* to remove cell debris. NMT activity was quantitated for each sample using an equal mass of supernatant protein as described.[27] Proteins separated by SDS–PAGE were transferred to nitrocellulose and probed from *Escherichia coli*.[28] Immunoreactive species were visualized with [^{125}I]protein A.

PWP1 DNA Sequencing

Overlapping fragments of the *PWP1* locus were subcloned from our original *NMT1* genomic clones[19] into the phagemid pRS316.[29] Single-strand DNA was purified from *E. coli* strain XL1-BLUE that had been transformed with the various subclones after infection with M13 helper phage R408 (Stratagene). The nucleotide sequences of both coding and noncoding strands were determined in their entirety by the dideoxy chain termination method.[30] Modified T7 DNA polymerase (U.S. Biochemical Corp.) plus "universal" M13 and T7 as well as synthetic oligonucleotide primers were used.

Construction of Plasmids and Mutant Alleles

Construction of the pBB110 plasmid and the *nmt1::HIS3* allele was previously described.[19] The three plasmids used to map the functional boundaries of *PWP1* (pBB147, pBB148, pBB149) were constructed by subcloning restriction fragments of the original *NMT1* genomic clones into the polylinker region of pRS316. The three subcloned fragments have a common endpoint at the *Hind*III site 3′ of *NMT1*, and therefore include the entire *NMT1* coding region (Fig. 1). pBB147 contains a 5.4-kb *Hind*III fragment, pBB148 contains a 4.3-kb Spe

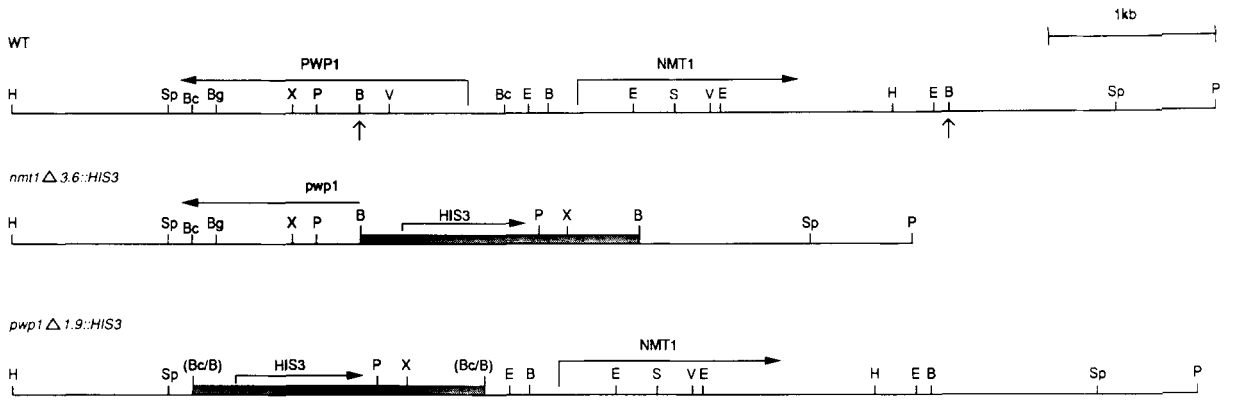

Fig. 1. Organization of the *NMT1–PWP1* locus and structure of deletetion mutants. All mutants were constructed as described in Materials and Methods, and their allelic designations appear above each map. Bent arrows represent protein-coding regions. The hatched box indicates the 1.7-kb *Bam*HI–HIS3 restriction fragment used to replace endogenous *NMT1–PWP1* genomic sequences: WT, wild-type; H, *Hind*III; Sp, *Spe*I; Bc, *Bcl*I; Bg, *Bgl*II; X, *Xho*I; P, *Pst*I; B, *Bam*HI; E, *Eco*RI; V, *Eco*RV; S, *Sph*I. Restriction sites appearing in parentheses have been destroyed during the cloning procedures.

I-*Hind*III fragment, and pBB149 contains a 3.65-kb *Xho*I–*Hind*III fragment (Fig. 1).

All genomic *nmt1::HIS3* and *pwp1::HIS3* alleles were made by transforming lithium acetate competent[31] YB100 with linear restriction fragments of DNA containing the desired mutation flanked by sequences homologous to the *NMT1–PWP1* locus, and selecting for histidine prototrophy.[32] Individual His$^+$ transformants were screened for the intended alterations by probing Southern blots of their genomic DNAs with ^{32}P-labeled probes extending across the *NMT1–PWP1* locus. To generate the *nmt1Δ3.6HIS3* allele (pBB137, Table I), the 3.6-kb region between the two *Bam*HI sites indicated by the arrows in Figure 1 was replaced with the *HIS3* gene. This was accomplished by cloning the 1.7-kb *Bam*HI*HIS3* fragment into the *Bam*HI site located between the 2.25-kb *Hind*III–*Bam*HI fragment 5′ of *NMT1* (Fig. 1) and the 1.6-kb *Bam*HI–*Pst*I fragment 3′ of *NMT1* (Fig. 1) present in a YIp5-derived plasmid. Cloning the 1.7-kb *Bam*HI*HIS3* fragment into *Bam*HI cut pBB157 created the *pwp1::HIS3* allele (pBB160). *Bcl*I digested pBB148 to generate the *pwp1Δ1.9::HIS3* (pBB167) allele (Table I).

Complementation of Mutant Alleles

All complementation tests were performed by transforming diploid cells heterozygous for the various mutant alleles with episomal plasmids, followed by sporulation and tetrad dissection.[31] Complementation was scored by rescue of the 2:2 (wild type:mutant) tetrads to 3:1 or 4:0 and by the presence of His$^+$, Ura$^+$ segregants (the *HIS3* gene was used to make all the mutants, and the *URA3* gene was present in all the complementing plasmids).

Computational Sequence Analyses

Database searching was carried out with the BLAST family of programs,[33] which are implemented as a network service at the National Center for Biotechnology Information (NCBI). The compute server for BLAST is a Silicon Graphics 4D/280, and the program is parallelized to use all eight processors simultaneously. Databases searched include: NBRF/PIR (release 27.0), SWISS-PROT (release 17.0), translated GenBank (release 66.0) and the GenInfo Backbone (GIBB, pre-release version Feb. 1991). More detailed descriptions of these data banks (and the BLAST program) can be found in reference 34. BLAST employs a new statistical theory for estimating the significance of local, ungapped sequence similarities.[35]

Sliding window comparisons and the calculation of z-value scores[36] were carried out using the RELATE program[37] on a VAX 8650 at the National Cancer Institute's Frederick Cancer Research Facility (NCI-FCRF). A window value of 27 residues was used and mean scores derived from 1,000 sequence randomizations were used to calculate z-values. Significant charge clusters[38] were identified using the Propat program generously provided by Volker Brendel of Stanford University.

The MACAW (Multiple Alignment Construction and Analysis Workbench) program,[39] running under Microsoft Windows on an IBM PS/2 model 70, was used to identify and estimate the significance of ungapped similarity blocks among the 11 nonorthologous β-transducin sequences (see Fig. 5). Once the most significantly related sequence common to all 11 proteins was identified, the MSA program[40] was used to compute the optimal alignment (with gaps) using a sequence weighting procedure[41] to control for data redundancy (see Fig. 6).

Prediction of secondary structure followed the "evolutionary comparison" strategy of Crawford et al.[42] Joint predictions for individual sequences were computed according to Nishikawa and Ooi[43] and then averaged over all the sequences for a joint consensus prediction (see Fig. 6).

TABLE I. Yeast Strains

Strain	Genotype
SK-1*	*MATa/MATα HO/HO* prototroph
YM2061**	*MATa ura3-52 his3Δ200 ade2-101 lys2-801 met⁻ LEU2::pRY181 (GAL1-LacZ) can1*
YB100**	*MATa/MATα ura3-52/ura3-52 his3Δ200/his3Δ200 ade2-101/ade2-101 lys2-801/lys2-801 met-/+tyr1-501/+ LEU2::pRY181/LEU2::pRY181 can1/can1*
YB102†	Isogenic with YB100 except *nmt1::HIS3/+*
YB151†	Isogenic with YB100 except *pwp1 nmt1Δ3.6::HIS3/++*
YB175†	Isogenic with YB100 except *pwp1::HIS3/+*
YB187†	Isogenic with YB100 except *pwp1Δ1.9::HIS3/+*
YB188‡	*MATa ura3-52 his3Δ200 ade2-101 lys2-801 met⁻ pwp1Δ1.9::HIS3 LEU2::pRY181 can1*

*See Johnston and Davis.[65]
**See Duronio et al.[18]
†Strain YB100 transformed with linear DNA carrying the indicated allele.
‡A segregant of strain YB187.

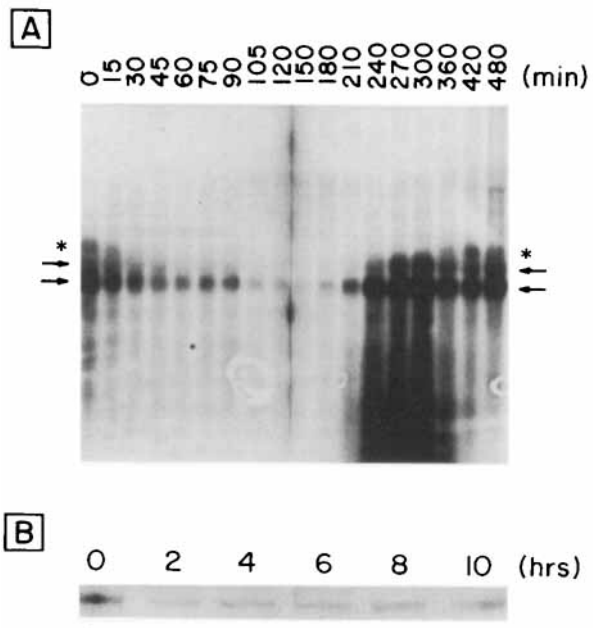

Fig. 2. Accumulation of *NMT1* mRNA during meiosis. **A.** Blot of total RNA, prepared from strain SK-1 at various times of meiosis and sporulation, was probed with a ^{32}P-labeled *Eco*RV–*Sph*I restriction fragment encompassing the 5′ ends of *PWP1* and *NMT1* mRNA (see Fig. 1). The lanes are designated according to the time (in min) RNA was harvested after shifting cells to SPM media. The star indicates the *PWP1* mRNA. The arrows note the two *NMT1* mRNAs. **B.** Western blot of crude lysates of sporulating SK-1 cells probed with rabbit polyclonal antiserum raised against *Escherichia coli*-derived *Saccharomyces cerevisiae* NMT. The lanes are designated as hours after shifting to SPM.

RESULTS AND DISCUSSION

Regulation of *NMT1* Expression

nmt⁻ cells do not arrest at the same stage of the cell cycle. This was determined by examining the morphological appearance of the small number of cells that resulted from germination of an *nmt1::HIS3* spore.[19] However, we wished to determine whether *NMT1* expression varied during mitosis or sporulation. Therefore, a blot of total RNA extracted at various times from a mitotically synchronized culture of cells (see Methods) was probed with a ^{32}P-labeled, 525-bp *Eco*RI restriction fragment derived exclusively from the coding region of *NMT1* (Fig. 1). Two *NMT1* transcripts were detected, and their levels remained constant over the course of the two cell cycles represented on the blot (data not shown). By contrast, the yeast POL30 transcript encoding the proliferating cell nuclear antigen subunit of DNA polymerase δ showed a cell cycle-dependent increase in steady state levels during the late G1 and S phase only.[44]

Diploid strain SK-1[45] is a homothallic prototroph useful for examining the temporal expression of genes during meiosis and spore formation: greater than 85% of the cells synchromously sporulate 30 h after shifting the cells from rich media to 1% potassium acetate (SPM). A blot of total RNA extracted from SK-1 at various times after switching to SPM was probed with a ^{32}P-labeled restriction fragment extending from the *Sph*I site in *NMT1* to and *Eco*RV site located 1,148 nucleotides upstream of the *NMT1* initiator methionine (Fig. 1). Figure 2 shows the two *NMT1* mRNAs that were detected. The smaller, more abundant transcript declined from its steady state concentration in log phase, vegetatively growing cells (Fig. 2A; $t=0$ min) to a level that was barely detectable at 150 min. Over the course of the next 150 min, the transcript increased to a concentration that was greater (150%) than the starting ($t=0$ min) level. At the latest time point examined ($t=480$ min), levels were 110% that of vegetatively growing cells. The larger *NMT1* transcript had a pattern of accumulation that mimicked that of the smaller transcript throughout meiosis (Fig. 2A). By contrast, the pattern of accumulation of yeast actin mRNA (*ACT1*) is similar to the two *NMT1* transcripts during the first 105 min of sporulation (i.e., gradually disappearing), but *ACT1* mRNA remains undetectable during the next 6 h.[44,*] These changes in NMT1 mRNA levels are not accompanied by changes in the steady state levels of the 53-kDa

*Note that the "mitotic" blots revealed that the ratio of the levels of the two *NMT1* transcripts were similar to that observed in the blots of RNA prepared from meiotically synchronized cells, although neither of their concentrations varied during the cell cycle.

acyltransferase (Fig. 2B). This is consistent with the notion that maternal *NMT1* mRNAs are deposited into newly forming spores.

A Newly Identified Gene, *PWP1*, Is Divergently Transcribed From *NMT1*

The data shown in Figure 2A revealed an additional transcript (indicated by an asterisk) that was not detected when the blot was probed with DNA derived only from the *NMT1* coding region (data not shown), suggesting the presence of a different gene closely linked to *NMT1*. A ^{32}P-labeled, 2.3-kb *Bam*HI–*Spe*I fragment located 5′ to *NMT1* (Fig. 1) also detected this RNA on the "mitotic blot" and indicated that its steady-state level did not vary during the cell cycle (data not shown). Since the temporal pattern of accumulation of this transcript (Fig. 2A) mimicked that of *NMT1* mRNA, we wondered whether the gene was transcribed from an overlapping or common promoter and whether its protein product affected N-myristoylation in yeast. We therefore characterized the genomic region upstream of *NMT1*.

The *nmt1::HIS3* recessive lethal allele can be fully complemented with an episome (pBB110) containing only *NMT1* and 185 nucleotides upstream of the *NMT1* initiator methionine.[19] By contrast, pBB110 does not fully complement a 3.6-kb deletion (*nmt1Δ3.6::HIS3*) of the *NMT1* locus that removes 1.15 kb of DNA upstream of *NMT1* (Fig. 1; Table I). When we attempted to rescue the *nmt1Δ3.6::HIS3* allele with the *NMT1* plasmid, an extremely slow growth phenotype resulted (see below). The functional boundary of the upstream gene could then be genetically mapped by attempting to fully complement the nmt1D3.6::HIS3 allele with plasmids containing different lengths of DNA 5′ of *NMT1* (see Methods). The results suggested that the boundary of the gene resided between the *Spe*I site and the *Xho*I site (Fig. 1).

Both strands of the 2.3-kb *Bam*HI–*Spe*I region 5′ to *NMT1* were sequenced. An open reading frame of 1,728 nucleotides was found that encoded a protein of 576 amino acids having a calculated M_r of 63,803 (Fig. 3). A termination codon was located 75 nucleotides upstream of the *Spe*I site, consistent with the complementation data. The gene is transcribed on the opposite DNA strand relative to *NMT1*. The predicted initiator methionine codons of the two proteins are separated by 664 nucleotides (Fig. 1).

Fig. 3. Nucleotide and predicted amino acid sequence of the *PWP1* gene. The *PWP1* coding strand is numbered beginning with the first nucleotide shown, and *PWP1* amino acids are numbered beginning with the predicted initiator methionine. The first eight residues of NMT are indicated at the top, and the bent arrow represents the 5′ end of one *NMT1* transcript previously determined by primer extension analysis.[19] Boxed nucleotides represents a canonical TATA box[68] for *PWP1*. Consensus sequences for transcription termination[69] are underlined. Pertinent restriction sites are also indicated (see text and Fig. 1). Tryptophan residues are underlined.

Fig. 3.

A

Query: PWP1 Periodic Tryptophan Protein - S. cerevisiae

Database: Protein Identification Resource 27.0 (complete), Dec. 31, 1990
26,798 sequences; 7,620,668 total residues.

SEQUENCES PRODUCING HIGH-SCORING SEGMENT PAIRS:	High Score	Smallest Poisson Probability P(N)	N
S09318 Nucleolar transcription factor UBF - Human	61	2.6e-05	4
A32265 Prothymosin alpha chain - Rat	52	5.2e-05	3
S01494 Hypothetical protein - Cowpox virus (fragment)	56	0.00012	4
S02708 Troponin T, skeletal muscle - Fruit fly	57	0.00078	2
S04631 Nucleolin - Human	52	0.0013	5
S08423 Ribosomal protein L12 - Haloarcula sp.	54	0.0034	2
A27272 Major centromere autoantigen CENP-B - Human (fragment)	59	0.0036	6
A28784 Sec7 protein - Yeast (Saccharomyces cerevisiae)	58	0.030	2
A29561 Prostatic spermine-binding protein precursor - Rat	66	0.037	2
A32230 Cytotactin precursor - Chicken	69	0.069	1
A30903 Protenascin precursor - Chicken	69	0.069	1
JQ0558 Flocculation suppressor protein SFL2 - Yeast	59	0.13	5
A29232 101K malaria antigen precursor - Plasmodium falcip...	66	0.15	2
TNBE11 91.8K alpha trans-inducing protein - Varicella-zost...	54	0.30	2
S00638 Regulatory protein LEU3 - Yeast (Saccharomyces cere...	52	0.37	3
S08415 Nucleolar protein NO38 - Chicken	52	0.38	2
JT0302 RNA polymerase delta subunit - Bacillus subtilis	62	0.53	1
H32354 Hypothetical protein (ctrA 5' region) - Bacillus su...	62	0.53	1
A26867 Cell division control protein 4 - Yeast (Saccharomy...	56	0.62	3

SIGNIFICANT LOCAL ALIGNMENTS:

>S09318 *Nucleolar transcription factor UBF - Human
Length = 764

Score = 61, Expect = 1.0, P-value = 0.65, length = 59

```
Query:    39 DDAKATLEEAEGESGVEDDAATGSSNKLKDQLDIDDDLKEYNLEEYDDEEIADNEGGKD 97
             DD     EE + E+G   + +  SS+   ++   D D +E + E+ DD+E  D++  ++
Sbjct:   686 DDEDEDEEEEDDENGDSSEDGGDSSESSSEDESEDGDENEEDDEDEDDDEDDDEDEDNE 744
```

Score = 55, Expect = 7.8, Poisson P-value = 0.0030, length = 60

```
Query:    50 GESGVEDDAATGSSNKLKDQLDIDDDLKEYNLEEYDDEEIADNEGGKDVSMFPGLSNDSD 109
             +E G +   +++  +    + + +DD  E + E+ D++E  ++EG+   S  +G S+DSD
Sbjct:   703 SEDGGDSSESSSEDESEDGDENEEDDEDEDDDEDDDEDEDNESEGSSSSSSSSGDSSDSD 762
```

Score = 50, Expect = 41., Poisson P-value = 0.00012, length = 18

```
Query:   557 DDQGEDEEEREGGDEHDD 574
             D++ EDE+E E  DE+ D
Sbjct:   684 DEDDEDEDEEEEDDENGD 701
```

Score = 46, Expect = 1.6e+02, Poisson P-value = 2.6e-05, length = 35

```
Query:   536 KSSRIARKYTSNDNPDTVITIDDQGEDEEEREGGD 570
             KSSR +    S+ + D     DD+ EDEEE +  +
Sbjct:   665 KSSRTTLQSKSESEEDDEEDEDDEDEDEEEEDDEN 699
```

Fig. 4A. Legend appears on page 48.

B

```
Query:  PWP1 Periodic Tryptophan Protein - S. cerevisiae (residues 150-524)

Database:  Protein Identification Resource 27.0 (complete), Dec. 31, 1990
           26,798 sequences; 7,620,668 total residues.
```

SEQUENCES PRODUCING HIGH-SCORING SEGMENT PAIRS:	High Score	Smallest Poisson Probability P(N)	N
A33928 GTP-binding protein beta chain homolog - Chicken	50	2.1e-05	5
B33928 GTP-binding protein beta chain homolog - Human	50	2.1e-05	5
JQ0558 Flocculation suppressor protein SFL2 - Yeast	59	0.054	3
A26867 Cell division control protein 4 - Yeast (Saccharomy...	56	0.072	3

SIGNIFICANT LOCAL ALIGNMENTS:

```
>A33928  GTP-binding protein beta chain homolog - Chicken
         Length = 317

  Score = 50, Expect = 17., Poisson P = 1.0, length = 16

Query:     31 DHTVKLWDLNSGNAAR 46
              D T++LWDL +G ++R
Sbjct:     84 DGTLRLWDLTTGTTTR 99

  Score = 39, Expect = 6.9e+02, Poisson P = 1.0, length = 14

Query:    215 MVIGGVNKVLKLWD 228
              +V G+ +K +KLW+
Sbjct:    120 IVSGSRDKTIKLWN 133

  Score = 39, Expect = 6.9e+02, Poisson P = 0.090, length = 18

Query:     26 ASTSADHTVKLWDLNSGN 43
              AS + D    LWDLN G+
Sbjct:    208 ASGGKDGQAMLWDLNEGK 225

  Score = 38, Expect = 9.7e+02, Poisson P = 0.0026, length = 14

Query:    224 LKLWDVFTNRSVRK 237
              L+LWD+ T  + R+
Sbjct:     87 LRLWDLTTGTTTRR 100

  Score = 38, Expect = 9.7e+02, Poisson P = 2.1e-05, length = 21

Query:    161 MSTGAMGEKTVKLWKFPLDDA 181
              M  +A  +KT+ +WK   D++
Sbjct:     30 MILSASRDKTIIMWKLTRDET 50
```

Fig. 4B. Legend appears on page 48.

C

Query: A33928 GTP-binding protein beta chain homolog - Chicken

Database: Protein Identification Resource 27.0 (complete), Dec. 31, 1990
26,798 sequences; 7,620,668 total residues.

SEQUENCES PRODUCING HIGH-SCORING SEGMENT PAIRS:	High Score	Smallest Poisson Probability P(N)	N
A33928 GTP-binding protein beta chain homolog - Chicken	1701	7.1e-246	1
B33928 GTP-binding protein beta chain homolog - Human	1701	7.1e-246	1
A26617 Transducin beta-2 chain - Bovine (fragment)	116	3.4e-14	5
B28040 Transducin beta-2 chain - Human	113	4.1e-14	5
B26617 Transducin beta-2 chain - Human	116	4.1e-14	5
JQ0558 Flocculation suppressor protein SFL2 - Yeast	98	1.2e-12	3
A30047 Enhancer of split protein - Fruit fly	86	2.1e-12	2
A32569 U4/U6 small nuclear ribonucleoprotein PRP4 - Yeast ...	93	4.2e-11	2
A32570 U4/U6 small nuclear ribonucleoprotein PRP4 - Yeast ...	93	4.2e-11	2
A30102 STE4 protein - Yeast (Saccharomyces cerevisiae)	65	1.9e-10	6
A24853 Transducin beta chain, liver - Human	114	2.1e-10	5
A24225 Transducin beta chain - Bovine	114	2.1e-10	5
A25457 Transducin beta chain - Bovine	114	2.1e-10	5
S05357 Hypothetical protein (clone AAC3) - Slime mold (Dic...	69	1.8e-09	6
A28040 Transducin beta-2 chain - Bovine (fragment)	116	2.7e-09	1
A26867 Cell division control protein 4 - Yeast (Saccharomy...	95	4.0e-09	6
A35096 Transducin beta-3 chain - Human	94	1.7e-08	4
A29938 MAK11 protein precursor - Yeast (Saccharomyces cer...	70	0.018	1

Fig. 4. Results of database searching. **A.** Complete 576-residue sequence of the *PWP1* gene product (Fig. 3) was used as a query to search the NBRF/PIR database using the BLASTP program with default parameters. Score refers to maximal segment pair (MSP) scores as defined in reference 35; High Score is the largest score when multiple MSPs were found in a single sequence, and N is the actual number of MSPs. The Expect values represent the calculated probabilities of finding similarities of this significance in the entire database; the Poisson values indicate the likelihood that the query sequence and the matched sequence are related by chance based on target frequencies of amino acids as described in reference 35. The + symbol used in the alignments indicates a conservative amino acid substitution. **B.** The *PWP1* amino acid sequence was edited to exclude subsequences spanning the N-terminal acidic domain and positive charge cluster, and the C-terminal acidic domain (see text) prior to repeating the database search as in A. **C.** Initial results of interative search strategy as described in the test. The top scoring sequence from the previous search (Chicken GTP-binding protein β-chain homolog, A33928) was used as a query sequence under the same conditions as the previous two searches. This procedure was repeated until the results converged on the set of twelve nonorthologous sequences presented in Table II. Orthologous sequences were excluded by the criteria of having a Karlin score of ≥1,000.

Sequence Analysis of *PWP1*

The results of initial database searches using the complete sequence of *PWP1* as the query, were dominated by significant local similarities to a number of proteins that contain sizeable acidic domains (Fig. 4A). Detailed inspection of the local alignments showed that *PWP1* contains two regions (approx. residues 23–109 and 524–576) that are highly enriched in aspartate and glutamate residues and that these two regions were largely responsible for the initial search results. Interestingly, many of the matched sequences were found in various nuclear proteins (Fig. 4A). However, because of the difficulties in interpreting the functional significance of matches that result from such locally "biased" amino acid compositions (see below), as well as the possibility that this phenomenon was obscuring other significant similarities, a new *PWP1* query was constructed to exclude the acidic domains. When this edited verison of *PWP1* was used as the query, highly significant ($p \sim 10^{-5}$) matches to several members of the β-transducin sequence family were observed (Fig. 4B). Intrasequence (self) comparisons were performed using standard computational methods[36] and revealed that, in addition to the acidic domains of *PWP1*, eight variable-length internal repeats containing an invariant tryptophan residue are present (Figs. 3, 4B, 5).

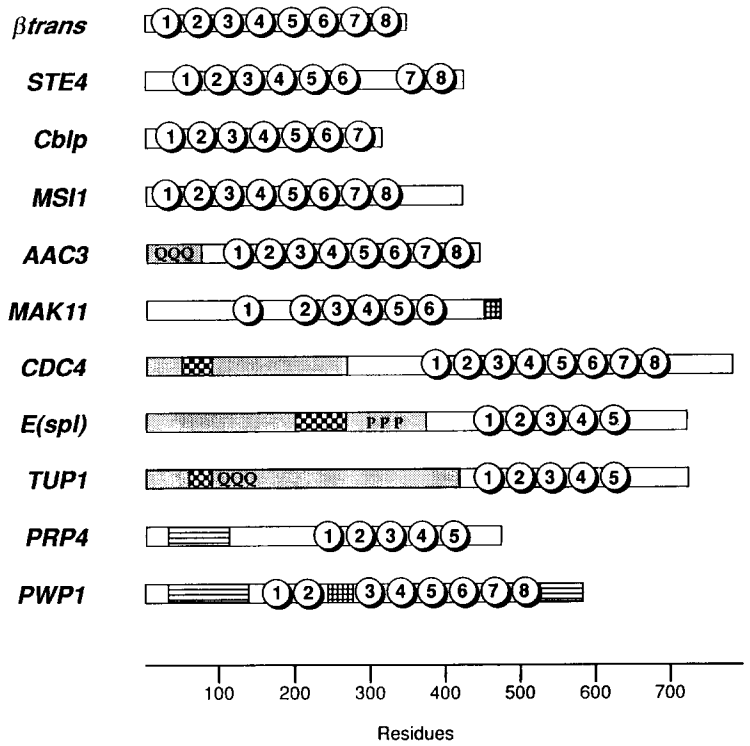

Fig. 5. Schematic diagram of β-transducin-related proteins. Rectangles represent linear sequences within the various proteins and are drawn approximately to scale. Numbered circles represent individual tryptophan-containing sequence repeats. Gray rectangles in AAC3, *CDC4*, *E(spl)*, and *TUP1* represent "low-entropy" subsequences as described in the text. The checkered boxes in *CDC4*, *E(spl)*, and *TUP1* indicate mixed charge clusters (see text). Horizontal lines denote acidic domains in *PRP4* and *PWP1*. The cross-hatched patterns in *MAK11* and *PWP1* mark the lysine-rich C-terminus and the central positive charge cluster (with nuclear targeting sequence), respectively. QQQ and PPP indicate oligo/poly-glutamine and proline-rich regions, respectively.

Comparative Analysis of β-Transducin-Like Proteins

Database searching

Although a number of reports on the identification of β-transducin homologs have appeared, no comprehensive analysis of this sequence family has been published; consequently, the functional significance of these homologies is often difficult to interpret. Thus, it was necessary to define the family quantitatively, which we did using both new and traditional methods.

First, iterative database searching[34] was carried out. From studies of protein families, it is known that not all pairwise relationships among family members are statistically significant, yet these relationships become clear and convincing when the entire network of sequence similarities is considered.[46] In other words, searching a database with a single query does not always reveal every related member of a sequence family. But if one iteratively takes as queries each significant match from the initial search, even the most distantly related members are eventually uncovered. So, for example, taking the chicken β-chain homolog (A33928, Fig. 4B) as a query sequence, we obtained the results shown in Figure 4C. Note the extremely high scores (and low *P* values) for the self-comparison. Of the nontrivial, nonredundant matches to the query sequence, five sequences (*SFL2*, a.k.a. *TUP1*), *Enhancer of split*, *PRP4*, *STE4*, and *CDC4*) had already been recognized as β-transducin homologs, but two new homologs were found: hypothetical protein AAC3 from *Dictyostelium* (S05357) and *S. cerevisiae MAK11* (A29938). Such iterative searches were carried out, and 11–12 (see below) nonorthologous β-transducin-like sequence families were found. Some of the properties of these sequences are summarized in Table II.

As in the case of the acidic domains of *PWP1*, certain regions of other β-transducin homologs (AAC3, *CDC4*, *E(spl)*, *TUP1*, and *PRP4*) created temporary difficulties in interpreting database search results. "Low-entropy" regions (see below) of these sequences are characterized by extremely skewed amino acid compositions, sometimes consisting of long runs of oligomers of a single residue, and database searches result in dozens to hundreds of spurious matches to other proteins with "biased" amino acid compositions (not shown). Such sequence attributes were actually exploited for classification purposes (see Discriminant analysis, p. 51), but

TABLE II. Summary of Biological Properties and Sequence Features*

Gene or protein name	Organism	Function and/or phenotype	Protein size in residues	No. of Trp repeats (location)	Other (computed) attributes	Database accession Nos.
β-Transducin (G_β)	Bovine retina	Component of heterotrimeric GTP-binding complex for transmembrane signal transduction	340	7–8 (entire seq.)		X03073 (GB) P11017 (SWP) A25457 (PIR)
STE4	*S. cerevisiae*	Component of signal pathway controlling response to mating pheromone; other components are *GPA1/SCG1* (G_α) and *STE18* (G_γ)	423	7–8 (entire seq.)		M23982 (GB) P18851 (SWP) A30102 (PIR)
Cblp	*C. reinhardtii*	Constituitively expressed during the cell cycle and flagellar regeneration	318	7 (entire seq.)		20116 (GIBB)
MHC-B 12.3 protein	Chicken liver	Associated with major histocompatibility complex in chickens but not in humans or mice	317	7 (entire seq.)		M24193 (GB) A33928 (PIR)
MSI1	*S. cerevisiae*	Probable component of cAMP pathway, negatively regulating RAS-mediated cAMP synthesis	422	8–9 (entire seq.)		M27300 (GB) P13712 (SWP) A36185 (PIR)
AAC3	*D. discoideum*	One of several developmentally regulated cDNAs containing repetitive AAC triplets (translates to polyglutamine in AAC3)	437	8 (central and C-term.)	Poly- and oligoglutamine (8–76)	X16524 (GB) P14197 (SWP) S05357 (PIR)
MAK11	*S. cerevisiae*	One of at least 32 *MAK* (maintenance of killer) genes necessary for M1 dsRNA replication; essential, membrane-associated	468	5–6 (central)	Low-entropy region; Lys-rich C-terminus; nuclear targeting seq. (452–456)	J03506 (GB) P20484 (SWP) A29938 (PIR)
CDC4	*S. cerevisiae*	Cell division cycle gene required early in mitosis; essential	779	8 (C-term.)	Mixed charge cluster 52–90); low-entropy region	X05625 (GB) P07834 (SWP) A26867 (PIR)
E(spl) (Groucho)	*D. melanogaster*	Segregation of neuroectoderm; interactions with other neurogenic loci (*Notch, Delta*) with homology to transmembrane mammalian growth factors; embryonic lethal phenotype	719	5–8 (C-term.)	Mixed charge cluster (199–272); proline-rich region (275–375)	M20571 (GB) P16371 (SWP) A30047 (PIR)
TUP1 (AER2, SFL2, CYC9)	*S. cerevisiae*	Pleiotropic phenotypes, including dTMP uptake, heme regulation and catabolite repression, *CYC7* expression, mating sterility, and plasmid stability	713	5–6 (C-term.)	Mixed charge cluster (66–89); oligoglutamine (101–118); low-entropy region	M35861 (GB) P18323 (SWP) JQ0558 (PIR)
PRP4 (RNA4)	*S. cerevisiae*	Component of U4/U6 snRNP with probable role in RNA splicing	465	5–8 (C-term.)	Negative charge cluster (83–108)	M26597 (GB) P20053 (SWP) A32569 (PIR)
PWP1	*S. cerevisiae*	Linkage and coordinate expression with *NMT1*; regulated during meiosis; nonessential	576	8–9 (dispersed)	Acidic domains; positive charge cluster (268–287); nuclear targeting seq. (273–277)	M37578 (GB) P21304 (SWP)

*Twelve nonorthologous members of the β-transducin sequence family were tentatively identified by iterative database searching as described in the text and legend to Figure 4. Subsequent analyses, however, indicated that the *Chlamydomonas* β-like polypeptide (Cblp)[11] and 12.3 protein found in chickens, humans, and mice[8] are actually orthologous sequences. In the case of some of the yeast genes, the most widely accepted name is used although synonyms are also provided in parentheses. The number of tryptophan-containing repeats cited in column 4 is generally those reported by the original investigators, although in some cases additional divergent copies are detectable using the most sensitive techniques. In cases where the repeats originally escaped detection, the number and location of repeat units were estimated from the sliding window comparisons (data not shown). The approximate locations of the repeats are indicated in column 4 and Figure 5 (more precise locations are given in the legend to Table III). Other computed attributes include low-entropy subsequences, charge clusters, and nuclear targeting motifs (column 5); in cases in which this column is blank, none of these characteristics were found. Regions of the sequences having low compositional complexity ("low entropy") were determined empirically by noting the locations and extents of spurious matches obtained when the complete sequences were used for database searching. Local biases in amino acid composition were estimated using the PRESIDUE program in PC-Gene (Intelligenetics, Inc.). More rigorous methods to assess low entropy sequences are under development (Wootton and Federhen, in preparation). Database accession numbers are given in the last column. Abbreviations used: GB (GenBank), SWP (SWISS-PROT), PIR (NBRF Protein Identification Resource), and GIBB (GenInfo Backbone).

TABLE III. Z-Value Scores Derived From Sliding Window Comparisons*

	BTRANS	*STE4*	*TUP1*	*PRP4*	CBLP	MHB123	*MAK11*	AAC3	*CDC4*	*E(spl)*	*MSI1*	*PWP1*
BTRANS	15.10	32.20	19.10	17.00	17.90	17.70	12.00	16.90	15.50	12.70	9.50	7.70
STE4		8.00	13.70	13.60	12.90	12.70	8.30	11.80	11.50	11.30	6.00	4.70
TUP1			18.80	22.40	20.80	19.90	15.10	21.30	24.90	17.40	6.80	13.30
PRP4				8.00	15.30	16.70	11.70	18.30	11.80	10.20	10.50	5.30
CBLP					13.40	58.50	11.50	14.50	17.80	16.50	6.20	10.20
MHB123						11.80	12.80	14.80	17.60	15.50	8.20	10.9
MAK11							5.90	9.70	11.00	8.70	5.60	6.20
AAC3								10.00	16.60	12.10	12.90	11.30
CDC4									11.00	11.60	8.10	9.80
E(spl)										7.40	4.80	7.20
MSI1											6.80	8.40
PWP1												5.30

*NBRF/PIR RELATE program was used to compute all pairwise and intrasequence (self-) comparisons for the eleven proteins presented in Table I. β-Transducin, *STE4*, Cblp, MHC-B 12.3, and *MSI1* are composed almost entirely of repeats, and their complete sequences were used. Because we wanted to restrict the comparisons to tryptophan repeat domains alone, other regions of the remaining sequences were excluded from the analyses. (These generally corresponded to the "low-entropy" subsequences described in Figure 5 and Table II, which, when included, produced artifactually high z-value scores (>50) that obscured other sequence features.) Thus, the following subsequences of these remaining proteins were used: AAC3 (residues 112–437), *MAK11* (residues 155–403), *CDC4* (residues 285–736), *E(spl)* (residues 439–695), *TUP1* (residues 407–669), *PRP4* (residues 176–465), and *PWP1* (residues 150–524). A window size of 27 residues was used, but the results were not significantly different were other windows (17 or 42 residues) were employed. The PAM250 matrix was used for scoring.

such regions must be excluded from statistical comparisons because they lead to extreme overestimation of significance and swamp other important sequence features.

Sliding window comparisons

Because repetitive-sequence proteins evolve by nonreciprocal recombination that often results in intrasequence transpositions, the relative order of internal repeats among paralogous proteins is often scrambled, making global sequence alignments (and scores based on them) difficult to interpret at best.[36] The traditional solution to this problem has been to compare such sequences using the "sliding window" approach.[36] The results of such an analysis for the β-transducin family are presented in Table III as a series of z-value scores.

β-Transducin itself is most closely related to *STE4* and most distantly related to *PWP1*. The former relationship is entirely congruent with the biochemical genetics of *STE4*, which is a known component of the mating pheromone signal transduction pathway in yeast whose other components, *GPA1* (or *SCG1*) and *STE18*, are strictly analogous to α-transducin and γ-transducin subunits, respectively.[7] Thus, z-value scores of the magnitude observed between β-transducin and *STE4* (32.2) may be considered indicative of *functional* similarity as well as sequence similarity.

Note also the extremely high score (58.5) between the *Chlamydomonas* β-like polypeptide and the C12.3 gene product from the chicken major histocompatibility complex. These two proteins are essentially identical in size and sequence alignments (not shown) demonstrate a uniform 74% sequence identity with only three small gaps. Even in the absence of supporting biochemical or genetic evidence, the conclusion here is that these sequences are *orthologous* and have identical functions in *Chlamydomonas*, chicken, man, and mouse.[8,11] The remainder of the pairwise scores are indicative of sequence homology (strictly speaking, descent from a common ancestor) but insufficient to infer functional analogy. We refer to such sequences as being *nonorthologous*; the term *paralogous* is also used.[36]

Intrasequence (self-) comparisons provide an estimate of how well conserved internal repeats are with respect to each other. Repetitive sequences within β transducin, *TUP1* and Cblp show the greatest degree of sequence conservation with z-values of about 13 to 18; repeats within *STE4*, *PRP4*, AAC3, and *CDC4* are moderately conserved with z-values in the range of 8–11; the most divergent internal repeats are found in *MAK11*, *E(spl)*, *MSI1*, and *PWP1* with z-values of 5–8 (Table III).

Discriminant analysis

Besides sequence similarity, other attributes of proteins (experimentally determined or computed) can be used for classification and functional prediction.[47] Known β-transducin homologs vary in size from 317 residues (Cblp) to nearly 800 residues (*CDC4*) and contain a variable number of tryptophan repeats (about 5–8). On this basis alone, it is difficult to conceive that all have basically the same function. β-Transducin, *STE4*, Cblp/MHC12.3 and *MSI1* are roughly equal in size and their sequences consist almost entirely of Trp-containing repeats (Table II; Fig. 5). Most of the remaining homologs appear to be mosaic proteins with Trp-repeats confined to specific domains.

In information theory,[48] the entropy of a message is a measure of its information content and in molecular sequence analysis it is increasingly common to refer to protein domains with skewed amino acid compositions as "low entropy" sequences.[36] For example, such low-entropy sequences are being increasingly found among transcription factors and proteins implicated in developmental regulation.[36] What this means, practically speaking, is that database searches employing such sequences as queries always result in a large number of matches in which no evolutionary homology or functional analogy can reasonably be inferred. Thus, the information content of such sequences, at least in terms of implying a specific biochemical function, is low.

More than one-half the known β-transducin homologs contain low-entropy regions apart from their Trp-repeat domains. Both AAC3 and *TUP1* contain poly– and oligoglutamine runs, whereas in *E(spl)* an extensive proline-rich domain is present. In other sequences, a small subset of amino acids may be overrepresented. In *CDC4,* for example, serine, alanine, asparagine, and threonine account for about 40% of the local amino acid composition in the shaded area (see Fig. 5). Like *PWP1,* the low-entropy region of *PRP4* is highly enriched for acidic residues.

Karlin and co-workers[38,49] suggested that rigorously defined charge clusters are useful attributes for classifying sequences and have applied this approach to a number of sequence families. We used their software to analyze the β-transducin homologs and found mixed (±) charge clusters in *CDC4, E(spl),* and *TUP1.* A positive charge cluster was identified in *PWP1*; this cluster encompasses a nuclear targeting signature:[50] Lys–Lys–Lys–Ser–Lys (residues 273–277). A negative charge cluster was found in *PRP4* corresponding to the previously mentioned acidic domain. Interestingly, an extended region of *TUP1* (residues 101–219) just distal to the mixed charge cluster was almost entirely devoid of charged residues.

In summary, biochemical and genetic evidence shows that STE4[51,52] and also probably *MSI1*[10] are functional analogs of β transducin and their sequence organizations and attributes are congruent with this biological context. The functions of Cblp and MHC12.3 are not known, but our studies strongly suggest that they also function in transmembrane signal transduction. AAC3 is very similar to these other sequences except for the N-terminal polyglutamine region. It is interesting to note in this context that the yeast γ subunit homolog, *STE18,* contains an N-terminal oligoglutamine region that apparently does not interfere with function.[7] Lastly, except for its C-terminal lysine-rich region, *MAK11* is also similar to this first group of sequences. In addition, *MAK11* appears to be a membrane-associated protein,[53] but its role in dsRNA replication is difficult to reconcile with transmembrane signal transduction.

CDC4, E(spl), and *TUP1* seem to form a related group. Each sequence contains a mixed charge cluster embedded in an N-terminal low entropy domain followed by a C-terminal Trp-repeat domain. *PRP4* and *PWP1* share some common characteristics: an N-terminal acidic domain followed by a largely C-terminal tandem Trp-repeat region. *PWP1* also has a central positive charge cluster and a C-terminal acidic region.

Alignment of Internal Repeats

For reasons previously stated (see Sliding Window Comparisons, above), we have not attempted to derive a global multiple alignment of tandem Trp-repeat domains among β-transducin-related proteins. Instead, we have taken a local alignment approach that is more tractable statistically[35] and has been embodied in the new tools for database searching[33] and multiple sequence alignment.[39] Using the MACAW program,[39] we identified the most significantly related subsequences common to all family members and an alignment of these is presented in Figure 6. Previous analysis of β transducin and CDC4 indicated that the average repeat size was about 41–43 residues.[4] Our results indicate that the core homology block is on the order of 24 residues and is centered around a Trp–Asp dipeptide (Fig. 6). Other residues outside of this core are also conserved, but one must invoke multiple insertion/deletion mutations to optimize the alignment. Perhaps these gapped regions, particularly the one found N-terminal to the core block, imply relaxed structural constraints (see below). Obviously, an analysis based on the most closely related repeat from each protein does not capture all the sequence variation represented in the family but serves as a detailed and firm reference against which other repeats and other proteins can be compared.

In several instances in which multiple sequences from evolutionarily diverse sources have been studied, secondary structure prediction has been accurate or useful.[42] The aligned, conserved repeats of β-transducin-related proteins suggest a series of β strands separated by turns (Fig. 6). This is in contrast to some other repetitive tryptophan proteins such as Myb proto-oncogene products[54] and TPR sequences,[55,56] which are thought to have predominantly helical conformations.

Possible Functions of Repetitive Tryptophans

Tryptophan occupies a unique place in protein structure and evolution because it is the least abundant and least mutable amino acid.[37] Therefore, tryptophan residues in β-transducin-related sequences are undoubtedly important for the functions of these proteins. What might the role of periodic tryptophans be? Multiple Trp-containing repeats

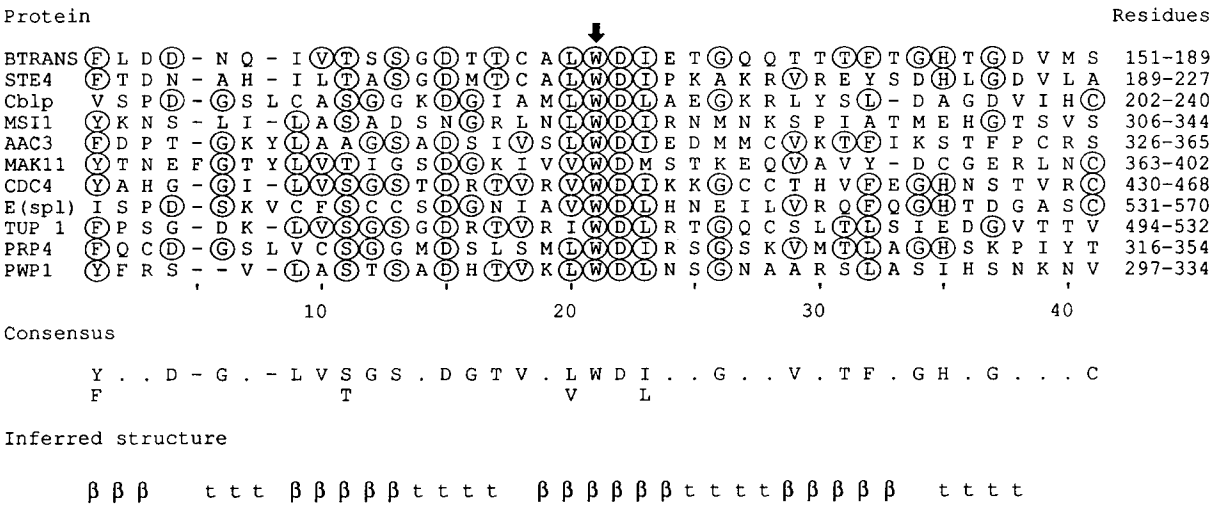

Fig. 6. Alignment of most highly conserved internal repeat. Complete sequences of the 11 nonorthologous β-transducin proteins (Fig. 5) were analyzed using the Multiple Sequence Alignment Construction and Analysis Workbench (MACAW) to locate the most significant region of local similarity common to all eleven sequences. This resulted in the identification of a 24-residue, ungapped homology block encompassing positions 9–32 in the displayed alignment. Because previous estimates placed the approximate size of an individual repeat unit at 43 residues,[4] the core homology blocks were extracted from their parent sequences along with 17 additional flanking residues and subjected to a full dynamic programming alignment permitting gaps.[40] Locations of these partial sequences within the parent proteins are given at the right margin. They correspond to the following repeat numbers as indicated in the schematic diagram (Fig. 5): 4, βtrans; 4, *STE4*; 5, Cblp; 6, *MSI1*; 7, AAC3; 5, MAK11; 4, CDC4; 3, *E(spl)*; 2, *TUP1*; 2, *PRP4*; 3, *PWP1*. Invariant and highly conserved residues are circled. The consensus sequence is based on a frequency of 33% and alternate residues are indicated if they are equally likely to be observed in a given position. The inferred secondary structure (β = beta strand, *t* = turn) was determined as described under Materials and Methods.

constitute the DNA binding domain of the Myb family of proteins, and it was suggested that Trp residues might be involved in "stacking" interactions.[57,58] Recent evidence from Raman spectroscopy indicates that these tryptophans are buried in the protein core; a model has been proposed in which three Trp residues form a cluster that stabilizes a helix-turn–helix DNA binding domain.[59]

PRP4 is a component of the small nuclear ribonucleoprotein particles (snRNP) that mediate splicing of nuclear pre-mRNA.[60,61] The *PRP4* protein is able to shield portions of U4 snRNA from hydrolysis by micrococcal nuclease[62]; analysis of U4 snRNA mutants indicates a specific RNA–protein interaction.[61] Because Myb-related and β-transducin-related proteins share no sequence similarity (apart from containing repetitive tryptophan residues), the helix-turn–helix model for Myb–DNA interaction is unlikely to be applicable to *PRP4*. A more relevant example perhaps is the large family of RNA binding proteins (termed RNP-CS[63]) in which repetitive aromatic residues may interact directly with nucleotide bases.[64] Photochemical cross-linking experiments showed that conserved, repetitive phenylalanine resides in the A1 hnRNP protein can form covalent adducts with oligodeoxynucleotides.[64] Perhaps the repetitive Trp residues in *PRP4* would have similar properties. Like *PRP4* (and *PWP1*), the RNP–CS family of RNA-binding proteins also possess acidic domains.[63]

Mutational Analysis of *PWP1*

As previously noted, *PWP1* contains two regions enriched in acidic amino acids and a striking positive charge cluster in addition to its Trp repeats (Table III; Fig. 5). Brendel and Karlin[38,49] proposed that highly charged regions may be discriminating features of regulatory proteins such as transcription factors. Several null mutations of *PWP1* were therefore generated to examine the role, if any, of this gene in cell viability, growth, and/or N-myristoylation. The *HIS3* gene was cloned into the *Bam*HI site at codon 220 of *PWP1* (see Fig. 3); a one-step gene disruption[32] was performed with diploid strain YB100 (Table I). After confirming the *pwp1::HIS3* heterozygous mutation by Southern blot hybridization analysis of genomic DNA (data not shown), the resulting YB175 strain (Table I) was sporulated and the tetrads dissected. The YB174 asci contained two wild-type spores and two spores that gave rise to extremely slow growing colonies. In all tetrads, the His^+ phenotype co-segregated with slow growth, indicating that the *pwp1::HIS3* gene disruption and the observed phenotype were tightly linked. In addition, centromere plasmids containing only the coding region of *PWP1* were able to fully complement the slow growth. To confirm the nonlethality of a *pwp1* null allele, nearly all *PWP1* coding sequences were deleted by replacing the 1.9-kb *Bcl*I restriction fragment with the *HIS3* gene. Tetrad analysis of

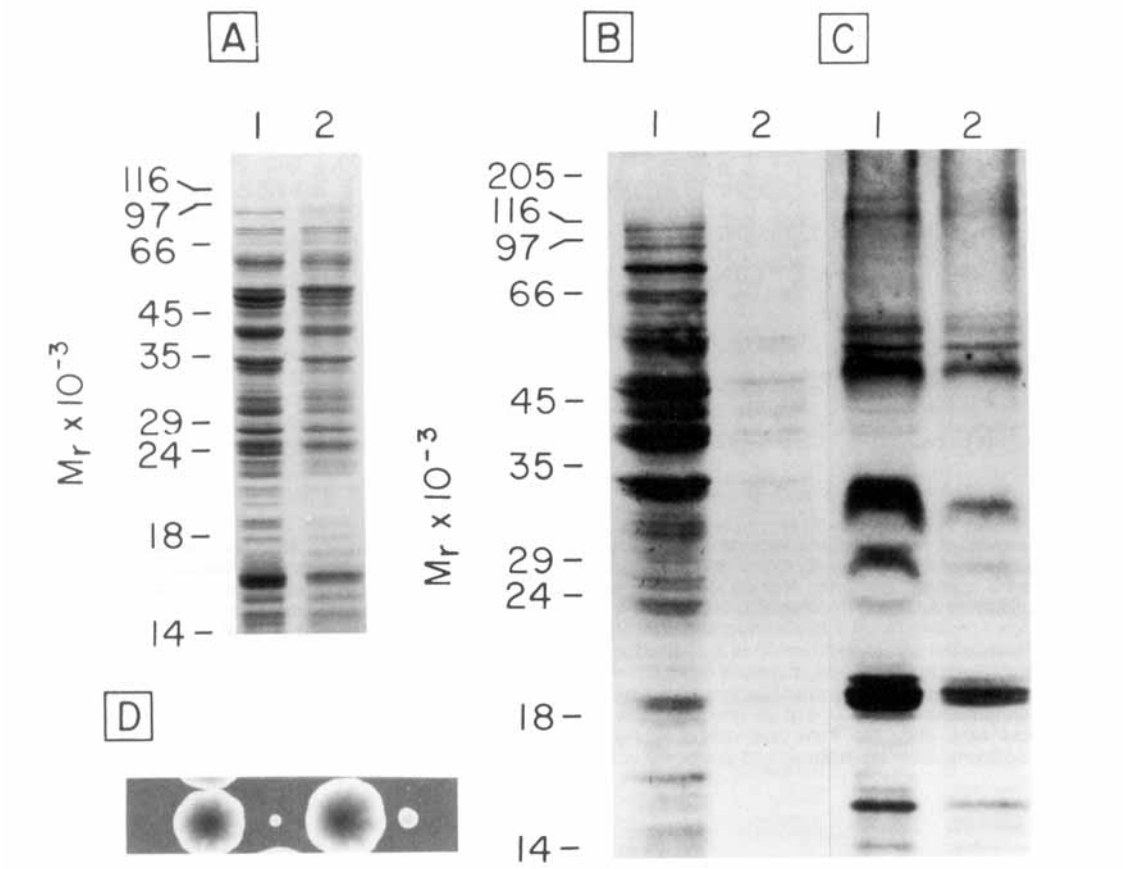

Fig. 7. Phenotype of the *PWP1* null mutant. **A.** Coomassie-stained SDS-polyacrylamide gel after electrophoretic separation of whole cell lysate proteins from strains YM2061 (wild type, **lane 1**) and YB188 (*pwp1Δ1.9::HIS3*, **lane 2**). (These strains were grown to identical densities in YPD prior to harvesting and lysing.) **B,C.** YM2061 cells (**lanes 1**) and YB188 cells (**lanes 2**) were grown to OD_{600} ~1.5 in YPD media and incubated for 1 h with either [^{35}S]methionine **(B)** or [^{3}H]myristic acid **(C)**. Metabolically labeled proteins in whole cell lysates were subjected to SDS–PAGE and fluorography. Note that the exposure time for B was 6 days and that for C was 45 days. **D.** Tetrad of strain YB187 (*pwp1/pwp1Δ1.9::HIS3*) dissected on a YPD plate and grown for 8 days at 30°C.

strain YB187 heterozygous for the *pwp1Δ1.9::HIS3* allele (Fig. 1, Table I) gave identical results to those obtained with YB175 (Fig. 7D).

Although the *pwp1* mutants grew very slowly, they remained viable. The growth rate of haploid strain YB188 (*pwp1Δ1.9::HIS3*, Table I) in liquid YPD culture was at least three times slower than that of a wild-type cell. To determine whether the attenuated growth reflected an alteration in protein N-myristoylation, wild-type YM2061 and YB188 cells were grown to an equivalent cell density and metabolically labeled with [^{32}S]methionine or [^{3}H]myristic acid for 1 h at 30°C. An equal amount of lysate proteins from each labeled culture was subjected to denaturing SDS–PAGE. The Coomassie-stained gels indicated that there was a pleiotropic effect on the steady-state level of several YB188 proteins (Fig. 7A). Incorporation of [^{35}S]methionine into strain YB188 proteins was also dramatically reduced compared to the wild-type strain (Fig. 7B). Labeling of cellular proteins with [^{3}H]myristate was generally not affected by deletion of PWP1, except for the absence of 33- and 29-kDa bands (Fig. 7C). Given the dramatic changes in [^{35}S]methionine labeling, this could reflect a failure to synthesize these particular proteins in *pwp1* mutants. Furthermore, the specific activity of NMT in YM2061 and YB188 was similar (data not shown).

Together these data suggest that the product of the *PWP1* gene does not serve a general function critical for protein N-myristoylation. The *S. cerevisiae* genome contains loci arranged in a similar manner that encode either gene products that functionally interact, e.g., the *GAL1–GAL10* locus[65] and the *TRT1* and *TRT2* gene pairs encoding histones H2A and H2B,[66] or gene products with no obvious biochemical relationship, e.g., the *PET56–HIS3* locus.[67] The observed phenotype of the PWP1 null alleles is consistent with the notion that this protein may have a regulatory role in cell growth and/or transcription. Its overall resemblance to *PRP4* (Fig. 5), its local similarities to acidic nuclear proteins (Fig. 4A), and its putative nuclear target motif are intriguing aspects for future study.

NOTE ADDED IN PROOF

Sethi et al. (Mol. Cell. Biol., in press) have cloned the yeast *CDC20* gene and shown that it is also a member of the β transducin superfamily. We agree with this conclusion but add that the sequence similarity to other known members is insufficient to make functional inferences and *CDC20* appears to represent yet another nonorthologous subfamily of G_β-related sequences. Sethi, N., Monteagudo, C., Koshland, D., Burke, D.J. (1991) The *CDC20* Gene of *Saccharomyces cerevisiae,* a β-transducin Homologue, is Required for a Subset of Microtubule-dependent Cellular Processes. Mol. Cell. Biol., in press.

ACKNOWLEDGMENTS

We thank Peter Burgers and Glen Bauer for supplying RNA blots and strain SK-1; Richard Zitomer for providing a manuscript prior to publication; and Alfred Gilman, Mark Goebl, Reed Wickner, and John Wootton for helpful discussions. This work was supported in part by grants from the National Institutes of Health (AI27179) and Monsanto Company. R.J.D. is the recipient of a predoctoral fellowship from the Spencer T. Olin Foundation.

REFERENCES

1. Gilman, A.G. G proteins: Transducers of receptor-generated signals. Annu. Rev. Biochem. 56:615–649, 1987.
2. Bourne, H.R., Sanders, D.A., McCormick, F. The GTPase superfamily: Conserved structure and molecular mechanism. Nature 349:117–127, 1991.
3. Reed, R.R. G Protein diversity and the regulation of signaling pathways. New Biol 2:957–960, 1990.
4. Fong, H.K.W., Hurley, J.B., Hopkins, R.S., Miake-Lye, R., Johnson, M.S., Doolittle, R.F., Simon, M.I. Repetitive segmental structure and the transducin β subunit: Homology with the *CDC4* gene and identification of related mRNAs. Proc. Natl. Acad. Sci. U.S.A. 83:2162–2166, 1986.
5. Yocum, J., Byers, B. Structural comparison of the yeast cell division cycle gene *CDC4* and a related pseudogene. J. Mol. Biol. 195:233–245, 1987.
6. Hartley, D.A., Preiss, A., Artavanis-Tsakonas, S. A deduced gene product from the *Drosophila* neurogenic locus, Enhancer of split, shows homology to mammalian G-protein β subunit. Cell 55:785–795, 1988.
7. Whiteway, M., Hougan, L., Dignard, D., Thomas, D.Y., Bell, L., Saari, G.C., Grant, F.J., O'Hara, P., MacKay, V.L. The *STE4* and *STE18* genes of yeast encode potential β and γ subunits of the mating factor receptor-coupled G protein. Cell 56:467–477, 1989.
8. Guillemot, F., Billault, A., Auffray, C. Physical linkage of a guanine nucleotide-binding protein-related gene to the chicken major histocompatibility complex. Proc. Natl. Acad. Sci. U.S.A. 86:4594–4598, 1989.
9. Dalrymple, M.A., Peterson-Bjorn, S., Friesen, J.D., Beggs, J.D. The product of the *PRP4* gene of *S. cerevisiae* shows homology to β subunits of G proteins. Cell 58:811–812, 1989.
10. Ruggieri, R., Tanaka, K., Nakafuku, M., Kaziro, Y., Toh-e, A. *MSI1*, a negative regulator of the RAS-cAMP pathway in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U.S.A. 86:8778–8782, 1989.
11. Schloss, J.A. A *Chlamydomonas* gene encodes a G protein β subunit-like polypeptide. Mol. Gen. Genet. 221:443–452, 1990.
12. Zhang, M., Rosenblum-Vos, L.S., Lowry, C.V., Aboayke, K., Zitomer, R.S. A yeast protein with homology to the β-subunit of G proteins is involved in control of heme-regulated and catabolite-repressed genes. Gene 97:153–161, 1991.
13. Towler, D.A., Gordon, J.I., Adams, S.P., Glaser, L. The biology and enzymology of eukaryotic protein acylation. Annu. Rev. Biochem. 57:69–99, 1988.
14. James, G., Olson, E.N. Fatty acylated proteins as components of intracellular signaling pathways. Biochemistry 29:2623–2634, 1990.
15. Heuckeroth, R.O., Glaser, L., Gordon, J.I. Heteroatom-substituted fatty acid analogs as substrates for N-myristoyltransferase: An approach for studying both the enzymology and function of protein acylation. Proc. Natl. Acad. Sci. U.S.A. 85:8795–8799, 1988.
16. Kishore, N., Lu, T., Knoll, L.J., Katoh, A., Rudnick, D.A., Mehta, P.P., Devadas, B., Huhn, M., Atwood, J.L., Adams, S.P., Gokel, G.W., Gordon, J.I. The substrate specificity of *Saccharomyces cerevisiae* Myristoyl-CoA:Protein N-Myristoyltransferase. J. Biol. Chem. 266:8835–8853, 1991.
17. Kobayashi, K., Suginaka, H., Yano, I. Analysis of fatty acid composition of *Candida* species by gas-liquid chromatography using a polar column. Microbios. 51:37–42, 1987.
18. Duronio, R.J., Rudnick, D.A., Johnson, R.L., Gordon, J.I. Myristic acid auxotrophy caused by mutation of *S. cerevisiae* myristoyl-CoA:protein N-myristoyltransferase. J. Cell Biol. 113:1313–1330, 1991.
19. Duronio, R.J., Towler, D.A., Heuckeroth, R.O., Gordon, J.I. Disruption of the yeast N-myristoyl transferase gene causes recessive lethality. Science 243:796–800, 1989.
20. Dietzel, C., Kurjan, J. The yeast SCG1 gene: A G_α-like protein implicated in the a- and a-factor response pathway. Cell 50:1001–1010, 1987.
21. Miyajima, I., Nakafuku, M., Nakayama, N., Brenner, C., Miyajima, K., Kaibuchi, K., Arai, K., Kaziro, Y., Matsumoto, K. *GPA1*, a haploid-specific essential gene, encodes a yeast homolog of mammalian G protein which may be involved in mating factor signal transduction. Cell 50: 1011–1019, 1987.
22. Jahng, K.-Y., Ferguson, J., Reed, S.I. Mutations in a gene encoding the α subunit of a *Saccharomyces cerevisiae* G protein indicate a role in mating pheromone signaling. Mol. Cell. Biol. 8:2484–2493, 1988.
23. Sewell, J.L., Kahn, R.A. Sequences of the bovine and yeast ADP-ribosylation factor and comparison to other GTP-binding proteins. Proc. Natl. Acad. Sci. U.S.A. 85:4620–4624, 1988.
24. Stearns, T., Willingham, M.C., Botstein, D., Kahn, R.A. ADP-ribosylation factor is functionally and physically associated with the golgi complex. Proc. Natl. Acad. Sci. U.S.A. 87:1238–1242, 1990.
25. Herman, P.K., Stack, J.H., DeModena, J.P., Emer, J.D. A novel protein kinase homolog essential for protein sorting to the yeast lysosome-like vacuole. Cell 64:425–437, 1991.
26. Bauer, G.A., Burgers, P.M.J. Molecular cloning, structure and expression of the yeast proliferating cell nuclear antigen gene. Nucl. Acids Res. 18:261–265, 1990.
27. Towler, D.A., Glaser, L. Protein fatty acid acylation: Enzymatic synthesis of an N-myristoylglycyl peptide. Proc. Natl. Acad. Sci. U.S.A. 83:2812–2816, 1986.
28. Rudnick, D.A., McWherter, C.A., Adams, S.P., Ropson, I.J., Duronio, R.J., Gordon, J.I. Structural and functional studies of *Saccharomyces cerevisiae* myristoyl-CoA:protein N-myristoyltransferase produced in *E. coli*. Evidence for an acyl–enzyme intermediate. J. Biol. Chem. 265:13370–13378, 1990.
29. Sikorski, R.S., Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. Genetics 122:19–27, 1989.
30. Sanger, F., Nicklen, S., Coulson, S.R. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74:5463–5467, 1977.
31. Sherman, F., Fink, G.R., Hicks, J.B. "Methods in Yeast Genetics." Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 1986.
32. Rothstein, R.J. One-step gene disruption in yeast. Methods Enzymol. 101:202–211, 1983.
33. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic Local Alignment Search Tool. J. Mol. Biol. 215: 403–410, 1990.
34. Boguski, M.S., States, D.J. In Chapter 5. "Protein Engi-

neering: A Practical Approach." Rees, A.R., Wetzel, R., Sternberg, M.J.E. (eds). Oxford: IRL Press, 1991.
35. Karlin, S., Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U.S.A. 87: 2264–2268, 1990.
36. States, D.J., Boguski, M.S. In Chapter 3. "Sequence Analysis Primer." Gribskov, M., Devereux, J. (eds). Salt Lake City: W.H. Freeman, 1991.
37. Dayhoff, M.O. "Atlas of Protein Sequence and Structure." Washington, D.C.: National Biomedical Research Foundation, 1978.
38. Brendel, V., Karlin, S. Association of charge clusters with functional domains of cellular transcription factors. Proc. Natl. Acad. Sci. U.S.A. 86:5698–5702, 1989.
39. Schuler, G.D., Altschul, S.F., Lipman, D.J. A workbench for multiple alignment construction and analysis. Proteins Struct. Func. Genet. 9:180–190, 1991.
40. Lipman, D.J., Altschul, S.F., Kececioglu, J.D. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 86:4412–4415, 1989.
41. Altschul, S.F., Carroll, R.J., Lipman, D.J. Weights for data related by a tree. J. Mol. Biol. 207:647–653, 1989.
42. Crawford, I.P., Niermann, T., Kirschner, K. Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase. Proteins 2:118–129, 1987.
43. Nishikawa, K., Ooi, T. Amino acid sequence homology applied to the prediction of protein secondary structure and joint prediction with existing methods. Biochim. Biophys. Acta 871:45–54, 1986.
44. Bauer, G.A., Burgers, P.M.J. Molecular cloning, structure and expression of the yeast proliferating cell nuclear antigen gene. Nucl. Acids Res. 18:261–265, 1990.
45. Johnston, L.H., Williamson, D.H., Johnson, A.L., Fennell, D.J. On the mechanism of premeiotic DNA synthesis in the yeast *Saccharomyces cerevisiae*. Exp. Cell Res. 141: 53–62, 1982.
46. Doolittle, R.F. Similar amino acids: Chance or common ancestry? Science 214:149–159, 1981.
47. Klein, P., Jacquez, J.A., Delisi, C. Prediction of protein function by discriminant analysis. Math. Biosci. 81:177–189, 1986.
48. Pierce, J.R. "An Introduction to Information Theory: Symbols, Signals and Noise." New York: Dover Publications, 1980.
49. Karlin, S., Blaisdell, B.E., Brendel, V. Identification of significant sequence patterns in proteins. Methods Enzymol. 183:388–402, 1990.
50. Bairoch, A. PROSITE: A dictionary of sites and patterns in proteins. Nucl. Acids Res. 19:2241–2245, 1991.
51. Whiteway, M., Hougan, L., Thomas, D.Y. Overexpression of the *STE4* gene leads to mating response in haploid *Saccharomyces cerevisiae*. Mol. Cell. Biol. 10:217–222, 1990.
52. Cole, G.M., Stone, D.E., Reed, S.I. Stoichiometry of G protein subunits affects the *Saccharomyces cerevisiae* mating pheromone signal transduction pathway. Mol. Cell. Biol. 10:510–517, 1990.
53. Icho, T., Wickner, R.B. The *MAK11* protein is essential for cell growth and replication of M double-stranded RNA and is apparently a membrane-associated protein. J. Biol. Chem. 263:1467–1475, 1988.
54. Kaneri-Ishii, C., Sarai, A., Sawazaki, T., Nakagosh, H., He, D.-N., Ogata, K., Nishimura, Y., Ishii, S. The tryptophan cluster: A hypothetical structure of the DNA-binding domain of the myb protooncogene product. J. Biol. Chem. 265:19990–19995, 1990.
55. Sikorski, R.S., Boguski, M.S., Goebl, M., Hieter, P. A repeating amino acid motif in *CDC23* defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. Cell 60:307–317, 1990.
56. Hirano, T., Kinoshita, N., Morikawa, K., Yanagida, M. Snap helix with knob and hole: Essential repeats in *S. pombe* nuclear protein nuc2+. Cell 60:319–328, 1990.
57. Anton, I.A., Frampton, J. Tryptophans in myb proteins. Nature 336:719, 1988.
58. Frampton, J., Leutz, A., Gibson, T.J., Graf, T. DNA-binding domain ancestry. Nature 342:134, 1989.
59. Kanei-Ishii, C., Sarai, A., Sawazaki, T., Nakagoshi, H., He, D.-N., Ogata, K., Nishimura, Y., Ishii, S. The tryptophan cluster: A hypothetical structure of the DNA-binding domain of the myb protooncogene product. J. Biol. Chem. 265:19990–19995, 1990.
60. Peterson-Bjorn, S., Soltyk, A., Beggs, J.D., Friesen, J.D. *PRP4 (RNA4)* from *Saccharomyces cerevisiae*: Its gene product is associated with the U4/U6 small nuclear ribonucleoprotein particle. Mol. Cell. Biol. 9:3698–3709, 1989.
61. Bordonne, R., Banroques, J., Abelson, J., Guthrie, C. Domains of yeast U4 spliceosomal RNA required for *PRP4* protein binding, snRNP-snRNP interactions, and pre-mRNA splicing in vivo. Genes Dev. 4:1185–1196, 1990.
62. Xu, Y., Peterson-Bjorn, S., Friesen, J.D. The PRP4 (RNA4) protein of *Saccharomyces cerevisiae* is associated with the 5′ portion of the U4 small nuclear RNA. Mol. Cell. Biol. 10:1217–1225, 1990.
63. Bandziulis, R.J., Swanson, M.S., Dreyfuss, G. RNA-binding proteins as developmental regulators. Genes Dev. 3:431–437, 1989.
64. Merrill, B.M., Stone, K.L., Cabioanchi, F., Wilson, S.H., Williams, K.R. Phenylalanines that are conserved among several RNA-binding proteins form part of a nucleic acid-binding pocket in the A1 heterogeneous nuclear ribonucleoprotein. J. Biol. Chem. 263:3307–3313, 1988.
65. Johnston, M., Davis, R. Sequences that regulate the divergent *GAL1–GAL10* promoter in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 4:1984.
66. Norris, D., Osley, M.A. The two gene pairs encoding H2A and H2B play different roles in the *Saccharomyces cerevisiae* life cycle. Mol. Cell. Biol. 7:3473–3481, 1987.
67. Struhl, K. Nucleotide sequence and transcriptional mapping of the yeast pet56–his3–ded1 gene region. Nucl. Acids Res. 13:8587–8601, 1985.
68. Hahn, S., Buratowski, S., Sharp, P.A., Guarente, L. Yeast TATA-binding factor protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. Proc. Natl. Acad. Sci. U.S.A. 86:5718–5722, 1989.
69. Zaret, K.S., Sherman, F. DNA sequence required for efficient transcription termination in yeast. Cell 28:563–573, 1982.