J Mol Evol (2000) 51:131–140
DOI: 10.1007/s002390010073

JOURNAL OF MOLECULAR EVOLUTION

Evolution of Simple Sequence in Proteins

Melanie Huntley, G. Brian Golding

Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada

Received: 13 January 2000 / Accepted: 17 April 2000

Abstract. The proteins of *Saccharomyces cerevisiae* contain a high proportion of low-complexity, simple sequences. These are protein segments composed almost exclusively or largely of a single repetitive amino acid polymer and are the most commonly shared feature between proteins. We have examined a survey of other species to determine how widespread this phenomenon might be. This was done by comparing how frequently segments from one protein are present in other proteins. Any recently evolutionarily related proteins were excluded. It was found that the most commonly shared features of eukaryotic proteins were repetitive but that prokaryotes did not contain such shared, extensively redundant repeats. The proportion of eukaryotic proteins that contain a significantly repetitive fraction changes dramatically from species to species. In addition the individual amino acids present in these repeats change between species. This suggests that the primary sequence of the repeats may not be important for their function. Further tests of the yeast repeats confirmed that these repeats evolve more quickly than the remainder of the protein sequence within which they are embedded. These results show that these rapid evolving, simple sequence repeats are in fact the most commonly shared pattern between all of the genomic proteins of eukaryotes.

Key words: Simple sequence — Low-complexity proteins — Amino acid repeats

Introduction

On a large scale, most proteins are composed of similar frequencies of the 20 amino acids. But with 20 residues possible at each site and most proteins composed of hundreds of amino acids, a random sequence of residues would still be unique. Functional constraints dictate that many proteins must accomplish similar tasks, such as binding to DNA or binding to ATP, and these functions are often accomplished by more or less distinct domains of amino acids. This leads to recognizable motifs that are now known to accomplish basic tasks (e.g., zinc finger motifs that help the protein bind DNA).

Another reason for similarity between proteins is their evolutionary origin. The most common origin of new proteins is thought to be from other proteins via gene duplication (Ohno 1987). After duplication of the gene sequence, the sequence can diverge to perform a new function. This can lead to similar domains in different proteins. Similar domains in distinct proteins can also be created via either homologous or nonhomologous recombination. Still another common suggestion is that these domains might have originated via exon shuffling (Gilbert 1978). Indeed, it has been suggested that an original advantage of introns was to facilitate the shuffling of exons (Dorit et al. 1990; de Souza et al. 1996; Gilbert et al. 1997). By whatever mechanism these domains were created, their existence has been well established in all organisms from bacteria to humans. The number of such domains, their frequencies, and many of their properties are less well known (Doolittle 1995).

Some proteins have unusual structures and may differ from these characteristic frequencies and motifs. Proteins such as unusually hydrophobic proteins or fibrous proteins (Creighton 1993) will have highly biased amino

Correspondence to: B. Golding

acid frequencies. In addition there are many proteins that have unusual amino acid compositions (Wootton 1994). These have been broadly termed repetitive sequences and became recognized when protein sequences were determined. The presence of repeated sequences within proteins has been detected in all organisms examined (Marcotte et al. 1998). However a survey of all the proteins in the *Saccharomyces cerevisiae* genome has revealed that a subclass of these repeats, repetitive simple sequences consisting mostly of one or a few amino acids, are the most common feature shared among yeast proteins (Golding 1999). In yeast a simple repetitive motif that is composed mostly of polyserine or related amino acids is most common. The next most common shared feature is a repeat composed mostly of polyglutamic acid and so on. The five amino acids S, E, D, Q, and N each compose the major residue within these repeats in *S. cerevisiae.*

Here the evolution of these repetitive structures has been examined in several organisms. Organisms whose genome has been completely sequenced were chosen to avoid the bias that easily soluble or otherwise "interesting" proteins would have and to determine their overall frequency within the entire genome. We show that these unusual repetitive simple sequences are a eukaryotic phenomena that does not appear to exist within bacteria or within archaebacteria despite the presence of "low-complexity" protein sequences within all of these organisms.

Materials and Methods

Six organisms that have been completely sequenced were chosen. The organisms chosen were: the Gram-negative bacteria *Escherichia coli* (Acc. No. U00096; Blattner et al. 1997), the Gram-positive bacteria *Bacillus subtilis* (EMBL No. BSUB9999; Kunst et al. 1997), the archaebacteria *Methanococcus jannaschii* (Acc. No. L77117; Bult et al. 1996), the archaebacteria *Pyrococcus horikoshii* (Acc. No. AP000001–AP000007; Tanaka et al. 1998), the eukaryote *Saccharomyces cerevisiae* (Goffeau et al. 1996, 1997; Mewes et al. 1997), and the eukaryote *Caenorhabditis elegans* (Acc. No. chr_I, chr_II, chr_III, chr_IV, chr_V, chr_X; The *C. elegans* Sequencing Consortium 1998). In addition to being completely sequenced, these organisms were chosen to cover a broad spectrum of the diversity of life. For each organism the protein sequences were collected from the public databases and analyzed separately.

The proteins, including translated open reading frames, from each organism were analyzed separately. Any redundant duplicates, isozymes or ancient duplications were removed. This was done by pairwise aligning different proteins from the genome and discarding the smaller of any two proteins that had greater than 20% identity. The percentage was calculated based on the total length of the alignment, and hence if a segment of amino acid sequence in one protein is similar to that in another and constitutes more than 20% of the protein, then the smaller of the two proteins will be excluded even if the two proteins are quite dissimilar in other parts of their sequence. A complete pairwise alignment was avoided by first screening for similar proteins in the genome using a BLAST search (Altschul et al. 1990) with a standard filter. All proteins that had a BLAST expect value less than 0.75 were then aligned using Clustal W (Thompson et al. 1994). This expect level was chosen to ensure that any potentially similar proteins would be aligned.

Each protein was divided into 100-residue-long segments. Segments that overlap by 80 residues were constructed. Segments were less than 100 amino acids if this composed the entire protein, and some segments were larger than 100 residues (but less than 120) at the COOH-terminus. This was done to maintain a constant 80-amino-acid overlap and an approximate 100-residue length. Each segment was then examined using the BLAST algorithm to determine how many other proteins contained a similar segment. The BLAST algorithm attempts to identify distantly evolutionarily related sequences and, in this case, compares protein residues using a PAM250 matrix. Protein sequences that contain subsequences that are more closely related in an evolutionary sense than expected by chance will have low expect values. The expect value itself is an estimate of how many proteins would be expected to be as closely related in a database of this size by chance alone. The database searched here consisted only of the nonredundant yeast proteins. For each segment, the number of BLAST hits with an expect value less than 0.05 were recorded.

To determine all potential simple protein sequence repeats it was, at times, necessary to eliminate the more frequent repeats to detect less frequent repeats. To accomplish this, the top 100-residue segment was examined to determine the most abundant amino acid residue. The protein segments were then screened against a 100-residue segment consisting solely of this one amino acid. All segments with a BLAST expect value less than or equal to 0.01 were removed. The remaining segments were then reexamined, and the next most frequent class of peptides was recorded. This process was repeated until the most frequent class of peptides showed no apparent pattern. In this way a series of classes of repeats is created, but it should be noted that these classes are not necessarily exclusive.

To discover if the repetitive simple sequences were evolving faster than the surrounding high complexity regions on a genome wide level, each of the 5,459 nonredundant yeast proteins were compared to all known proteins in the public databases. The most closely related sequence (according to a BLAST criterion) that did not belong to *S. cerevisiae* for each protein was collected. Clustal W was used to align these two; the yeast sequence and its most similar nonyeast homologue. Only 3,253 sequences had a BLAST expected value less than 1.0×10^{-5} and were considered further. For each of these pairwise alignments, the 100-residue region that contained the most frequent repeat was identified. Then the percentage of base substitutions and the percentage of indels were compared inside and outside of this repetitive 100-residue region. The significance of this difference was tested using a standard Z-score test of the difference between two percentages.

Results

The elimination of closely related proteins considerably reduces the number of distinct proteins considered from each organism. The number of proteins considered for each species is shown in Table 1. As an example, the genome of *C. elegans* contains 17,083 putative proteins. These were reduced to just 9,685, a 43% reduction. This table also shows the amino acid composition of these proteins. Between species there are subtle differences in amino acid content, but there are not apparent dramatic changes. The relative frequencies of amino acids between species are roughly comparable.

Genome Content

The two most frequent segments shared between distinct proteins are shown in Table 2. This table gives the num-

Table 1. The frequency of amino acids among nonredundant proteins

Amino acid	*Escherichia coli*	*Bacillus subtilis*	*Methanococcus jannaschii*	*Pyrococcus horikoshii*	*Saccharomyces cerevisiae*	*Caenorhabditis elegans*
A	0.094	0.074	0.056	0.062	0.053	0.062
C	0.012	0.008	0.013	0.007	0.013	0.020
D	0.053	0.053	0.055	0.043	0.058	0.054
E	0.059	0.074	0.086	0.084	0.066	0.068
F	0.038	0.044	0.042	0.046	0.046	0.046
G	0.072	0.067	0.064	0.069	0.048	0.051
H	0.023	0.023	0.014	0.015	0.022	0.024
I	0.059	0.072	0.104	0.087	0.066	0.060
K	0.046	0.073	0.103	0.078	0.074	0.065
L	0.105	0.095	0.093	0.102	0.098	0.085
M	0.028	0.028	0.023	0.024	0.021	0.026
N	0.040	0.041	0.052	0.036	0.062	0.050
P	0.045	0.036	0.034	0.046	0.043	0.050
Q	0.045	0.039	0.015	0.017	0.039	0.042
R	0.056	0.042	0.039	0.055	0.045	0.054
S	0.058	0.063	0.045	0.058	0.091	0.083
T	0.054	0.054	0.041	0.045	0.058	0.058
V	0.069	0.067	0.069	0.076	0.055	0.061
W	0.015	0.010	0.008	0.012	0.010	0.010
Y	0.029	0.036	0.044	0.039	0.033	0.030
Total amino acids	887,389	779,407	359,773	420,549	2,266,886	4,475,410
Total distinct proteins	2,898	2,790	1,266	1,527	5,459	9,685

ber of other proteins from the same organism that have a similar segment according to the BLAST criterion. The most frequent segment for *E. coli* contains 38 alanine residues, 30 serine residues, 11 threonine residues, and 21 others. It has similar peptides in 32 out of the 2,898 distinct proteins, comprising 1% of all *E. coli* proteins. The protein containing this segment is b1372, a putative membrane protein. The most frequent 100-residue segment for *B. subtilis* contains 26 glutamic acid residues, 21 lysine residues, 11 aspartic acid residues, and 53 others. This segment has significant similarity to 1% (31 of 2,790) of the genomic proteins. The protein that contains this segment is yukC, whose function is unknown.

The archaebacterial genomic proteins have segments with similarity to a somewhat greater percentage of the proteins. The most frequent segment from *M. jannaschii* contains 22 lysine residues, 15 glutamic acid residues, 14 leucine residues, and 49 others. This segment has significant similarity to 4% (54 of 1,266) proteins. The protein that contains this segment is MJ1254, a hypothetical protein. The most frequent segment from *P. horikoshii* contains 32 glutamic acid residues, 20 lysine residues, 11 leucine residues, and 37 others. This segment has a significant match in 4% (57 of 1,527) of the genomic proteins. The protein that contains this peptide is PH0553, a hypothetical protein.

The eukaryotic genomic proteins have far greater similarity from protein to protein and greater simplicity. The most frequent segment from *S. cerevisiae* is rich in serine residues, with 51/100 of them. It also contains 27 glutamic acid residues, 10 lysine residues, and 12 other residues. This peptide segment has similar segments in 14% of all yeast proteins (754 out of 5,459 proteins). The protein containing this peptide is SW-NSR1, a nuclear localization sequence binding protein. The most frequent segment from *C. elegans* is rich in glutamic acid residues, with 51 of them. It also has 33 lysine residues, 6 aspartic acid residues, and 15 others. Similar peptide segments occur in 8% of the proteins (763 out of 9,685). The protein containing this peptide is g3875441, a hypothetical protein.

Although their genomes are not yet completely sequenced (at the time of this analysis), we also collected 1,000 proteins from *A. thaliana* and 478 from *D. discoideum*. Although *A. thaliana* has repeats in its genomic proteins, they are not as extensive or as common as those found in other eukaryotes. The most common is a mixture of glutamic/aspartic acid and arginine and is shared by 5.4% of these 1,000 proteins. Other prevalent repetitive simple sequences in *A. thaliana* included polyproline. On the other hand, repetitive simple sequences were exceptionally common among 478 unrelated proteins from *D. discoideum.* The common presence of repeats in some *D. discoideum* proteins has been previously observed by Shaw et al. (1989). The most common repeat is polyasparagine shared among 32% of the distinct proteins.

The Gram-negative and Gram-positive bacteria and the two archaebacteria showed similar results. For all four organisms representing two domains of life, the segment with the highest number of similar segments in other proteins was not noticeably repetitive or rich in any particular amino acid residue (Table 2). Although they had similar patterns, the percentage of proteins with pep-

Table 2. Most common nonoverlapping protein segments

Count	Protein	Sequence	Identification
Escherichia coli			
32	bl372	KTAAASSASAASTSAGQASASATAAGKSAESAASSASTATTKAGEATEQA SAAARSASAAKTSETNAKASETSAESSKTAAASSASSAASSASSASASKD	Putative membrane spanning domain
30	tolA	AAADAKAKAEADAKAAEEAAKKAAADAKKKAEAEAAKAAAEAQKKAEAAA AALKKKAEAAEAAAAEARKKAATEAAEKAKAEAEKKAAAEKAAADKKAAA	Putative membrane spanning domain
Bacillus subtilis			
31	yukC	NDYIYFALAKYKQQLLSEDTNDEDIQKELDSVNSELEKAQKERQENKQSN SETSLVDTSEEQTQTDEEKQAEEKAAEEKAAAEEKAKKEEQKEKEDEKKE TEKKDEKKDDK	Unknown function
28	yttA	KEHEELEKEYKSVSSEAKKLKDNKEDQDKLEKLKNENSDLKKTQKSLKAE IKELQENQKQLKEDAKTAKAENETLRQDKTKLENQLKETESQTASSHEDT	Unknown function
Methanococcus jannaschii			
54	MJl254	RTKNIKNELTSLKNKLKEKEEEIKNLAIKIKDLEDKLSKANKNLLNKDEI ISVLNERISEYESQIQKLLDENIIYKEKIESLNKYIETLKKENDKLKDKV	Predicted coding region
54	MJ0884	VAADLVDYFEIKEDELKVLVGDKLASEILKILKEKKKLERKKKKEKEKLE KEKKKEEKAKEKQSNLIIQPKEIKEEVKAEVEKKEEVKEKIVEKPKAEEV	Putative activator (replication factor C)
Pyrococcus horikoshii			
57	PH0553	LSSQLSRLVEALEEKKFAVHEKKAESIAEKAAEVTEKVERIEELLEEKPK EEKSELAKKVEEIHKKVEELEEKLTGEKLEETKKKVEELEEKIEKGEEVT	Hypothetical protein
33	PH1798	LRIRMSDVEKEISLISKDLEKLIKEEESLRSEIEDSERKIAEIDETISKK KDEVAKLKGRIERLEKRRDKLKKALENPEAREVTEKIREVEREIAKLREE	Hypothetical chromosome assembly protein
Saccharomyces cerevisiae			
754	SW-NSR1	SSSESESESESESESSSSSSSSDSESSSSSSSDSESEAETKKEESKDSSS SSSDSSSDEEEEEEKEETKKEESKESSSSDSSSSSSSDSESEKEESNDKK	Nuclear localization sequence binding protein
745	SW-SR40	SSSSSSSSSGESSSSSSSSSSSSSSDSSDSSDSESSSSSSSSSSSSSSSS DSESSSESDSSSSGSSSSSSSSSDESSSESESEDETKKRARESDNEDAKE	Putative supressor protein srp40
Caenorhabditis elegans			
763	g3875441	AEKNEEDKKEEEPKKEEEKKEEVEKKEEDEKKDEEPKKEEEKKEEEQKEE VEKKEEEEKKDEEPKKEEEKKEEEEKKEDEVEEKSEKVEEKELEPKKDEE ETKKN	Hypothetical protein
733	g3523105	KSRKRAKSESESDESDEEEDRKKSKSKKKVDQKKKEKSKKKRTTSSSEDE DSDEEREQKSKKKSKKTKKQTSSESSEESEEERKVKKSKKNKEKSVKKRA	Similar to SNF2/RAD54 family of helicases
Arabidopsis thaliana			
54	AC007369_9	DLKKSRRDRDRSNERKKDKGSEKRREKDRRKKRVKSSDSEDDYDRDDDEE REKRKEKERERRRRDKDRVKRRSERRKSSDSEDDVEEEDERDKRRVNEKE	Similar to RNA helicases
47	CAB43856	KQRKCISEKKPLKKPEVSTDEEEEEEEENEQSDEGSESGSDLFSDGDEEGN NDSDDDDDDDDDDDDDDEDAEPLAEDFLDGSDNEEVTMGSDLDSDSGGSK	Putative protein
Dictyostelium discoideum			
154	P3K2_DICDI	CNNLTSSSSSSSTTATTPSPTTTSNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNTTSTTTTTTSILISSSPPPSSSSSSSSSNDEQFNN	Phosphatidylinositol 3-kinase 2
153	CAA71241	SQQDLSTISSPILSSSTTSSSSSISTDSNLSSNNNNNNNNNNNNSTPILS STSTTTTTTTTNNNNNSNNTFQPISVSKSSSFSKSTISTNPSSKSSSNL	racGAP (racGTPase-activating protein)

tides similar to the most common peptide segment was 4% in both the archaebacteria, and only 1% in the Gram-positive bacteria. These percentages are in striking contrast to the repetitive richness of the serine residues in *S. cerevisiae* and the glutamic acid residues in *C. elegans*. The most common segment in *E. coli* has similar peptides in 1% of all of *E. coli*'s proteins, but *S. cerevisiae* and *C. elegans*'s common segments have similar peptides in 14% and 8% of their proteins respectively.

Another feature that distinguishes the eukaryotes from the prokaryotes is the presence of multiple classes of repetitive sequences. Table 3 shows that both of the finished eukaryotic genomes had at least five distinct repetitive classes each (as defined in Materials and Methods). In *S. cerevisiae* these consist of the five amino acids S, E, D, Q, and N. In *C. elegans* the amino acids composing the most common segments are E, T, K, S, Q, P, R, and G. Although these lists of amino acids from each organism are similar (each containing S, E, and Q), there are repeats shared between many genomic proteins in one organism that are composed of an amino acid that does not appear to be present in excess in the other organism. In addition to those amino acids that are shared, their relative order in terms of frequency has changed.

Both *C. elegans* and *S. cerevisiae* have a large number of glutamic acid repeats (class "b" for *S. cerevisiae* and class "a" for *C. elegans* in Table 3). But this does not skew the overall, genome wide concentration of glutamic acid. All of the prokaryotic species except *E. coli* have higher frequencies of glutamic acid than do these two eukaryotes (Table 1).

Table 3. The frequency of the most common classes of 100 mer's

Class	Frequency	Protein	Representative sequence
Saccharomyces cerevisiae			
a	13.8%	SW-NSR1	SSSESESESESESESESSSSSSDSESSSSSSSDSESEAETKKEESKDSSS SSSDSSSDEEEEEEKEETKKEESKESSSSDSSSSSSSDSESEKEESNDKK
b	7.8%	SW-YKU1	KKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKQEEEEKKKKEEEEKKKQ EEGEKMKNEDEENKKNEDEEKKKNEEEEKKKQEEKNKKNEDEEKKKQEEE
c	4.4%	SW-YGG6	CGKPLALTAIVDHLENHCAGASGKSSTDPRDESTRETIRNGVESTGRNNN DDDNSNDNNNDDDDDDDNDDNEDDDDADDDDDNSNGANYKKNDSSFNPLK
d	3.8%	SW-ADR6	NNNNSNNHNMRNNSNNKTSNNNNVTAVPAATPANTNNSTSNANTVFSERA AMFAALQQKQQQRFQALQQQQQQQQNQQPQQQQQQQQNPKFLQSQ
e	3.6%	PIR-S61046	NMAPSNSGSPIIIADHFSGNNNIAPNYRYNNNINNNNNNINNMTNNRYNI NNNINGNGNGNGNNSNNNNNHNNNHNNNHHNGSINSNSNTNNNNNNNNGN
Caenorhabditis elegans			
a	7.9%	g3875441	AEKNEEDKKEEEPKKEEEKKEEVEKKEEDEKKDEEPKKEEEKKEEEQKEE VEKKEEEEKKDEEPKKEEEKKEEEEKKEDEVEEKSEKVEEKELEPKKDEE ETKKN
b	5.8%	g4226144	TTTTSTTSSTTTTTATSTTESTSTSTDSTTTESTTESTTESTSTSTDSTT TESTTESTTESTSTSTDSTTTESTSTSTDSTTTESTTESTTESTSTSTDS
c	5.7%	g291182	DSSDNSDSSESSDDDKKSKKKKKKSKKDSSDSSDSSDSSDSSDDGKKKKK KKSKKDSSDDSSDSSDSSDSSDSSDDDKKKKKKSKKSKNKGSSSDSSDSS
d	4.9%	g3877936	ARKKKEKTPTPTESSFESSSDSSSTSESSTSSESSSSASESESESKSESQ VSSSKTSTSKASSSKAYGSDFESEKSSSSSASTISKVTPKKLDKPQKTKK
e	3.1%	g3879933	QQQQQQREQQQREQQHREHQARLQQHQQQQQQQQQQQQQRPPQPQPQPQP QPPQRPPQQPQSFSGTHELHLQRQREQQQQQQQQQQQQQQRQQNPQQQPQ
f	3.0%	g3874621	PKTPPPPPPMQHQNHQNHQYQQQHPSLPRSASTPQPIQQQQSSIPPPPP PPPPHCEPTMVHVEFTPPSTSSVPPPPPPLPPISSGAPPPPPPPPPGGLM
g	2.0%	g388602	MSRRSRSRSRSPKRDREERKRREDRDRDRERKRDRKDRERKRRHRSSSSE GSQAEPHQLGSIFREERRRRERNESPKLPPPPPPPPSDPPVDTSIPFDVS
h	1.8%	g3875269	PQPASCGCAPACPQAPSCPVCPPPQPCPAPPAAYCPQVQPVYQSGGGGC GGGGCGGGGGGCGGGGGCGGGGGGGCGGGGGGGCGGGGGGGGGGYASGGS

Relative Evolutionary Rate

Two examples of yeast proteins aligned with homologues that have a BLAST expect value less than 1.0 × 10^{-5} were chosen. Both of these examples were chosen on the basis that the Clustal W output showed good similarity in the regions directly adjacent to the repeat and that the proteins appeared to have similar descriptions in each organism.

The first example is a 705-amino-acid heat shock cognate protein, hs83 (Borkovich et al. 1989; Accession P15108). The 100-residue segment from this protein is similar to sequences in 80 other distinct yeast proteins. The 18 BLAST hits for this protein include AAA02813 from *S. cerevisiae,* P46598 from *Candida albicans,* AAB97626 from *Podospora anserina*, P41887 from *Schizosaccharomyces pombe,* A48426 from *Zea mays,* Q08277 from *Z. mays*, P54651 from *D. discoideum,* P06660 from *Trypanosoma cruzi,* A26125 from *T. cruzi,* CAA99793 from *C. elegans,* P02829 from *S. cerevisiae,* S57415 from *Leishmania donovani infantum*, AAD41357 from *Tetrahymena thermophila*, A44983 from *Trypanosoma brucei*, P27741 from *Leishmania amazonensis,* P12861 from *Trypanosoma brucei brucei*, AAA66179 from *Plasmodium falciparum,* AAB35313 from *Leishmania braziliensis.* From the 6th to 218th amino acid all the sequences (except the *L. braziliensis* sequence, which doesn't begin until the 49th amino acid), align with great similarity (results not shown). Figure 1 shows the alignment around a glutamic acid and lysine repeat in the yeast sequence. Although the other sequences are also rich in these two residues in this area, the similarity is weaker. At amino acid 259, where the repeat region ends, the sequences align extremely well once again. This strong similarity continues to the end of the proteins. The repetitive simple sequence in the middle of the protein shows high levels of indel substitutions in comparison to the closely related sequences.

The second example is a 1,235-amino-acid high-affinity potassium transport protein, trk1 (Gaber et al. 1988; Accession P12685). The 100-residue segment from this protein is similar to sequences in 133 other distinct yeast proteins. The 12 proteins with BLAST expect values less than 1.0 × 10^{-5} are P28569 from *Saccharomyces bayanus,* P28584 from *S. cerevisiae,* CAA08813 from *Neurospora crassa,* AAD30128 from *Kluyveromyces lactis,* Q10065 from *S. pombe,* S50225 from *S. pombe*, P47946 from *S. pombe*, CAB39784 from *A. thaliana,* AAC62807 from *A. thaliana*, S47582 from *Triticum aestivum,* BAA18016 from *Synechocystis* sp., and AAC07434 from *Aquifex aeolicus.* The first eight sequences align with good similarity from amino acid 56

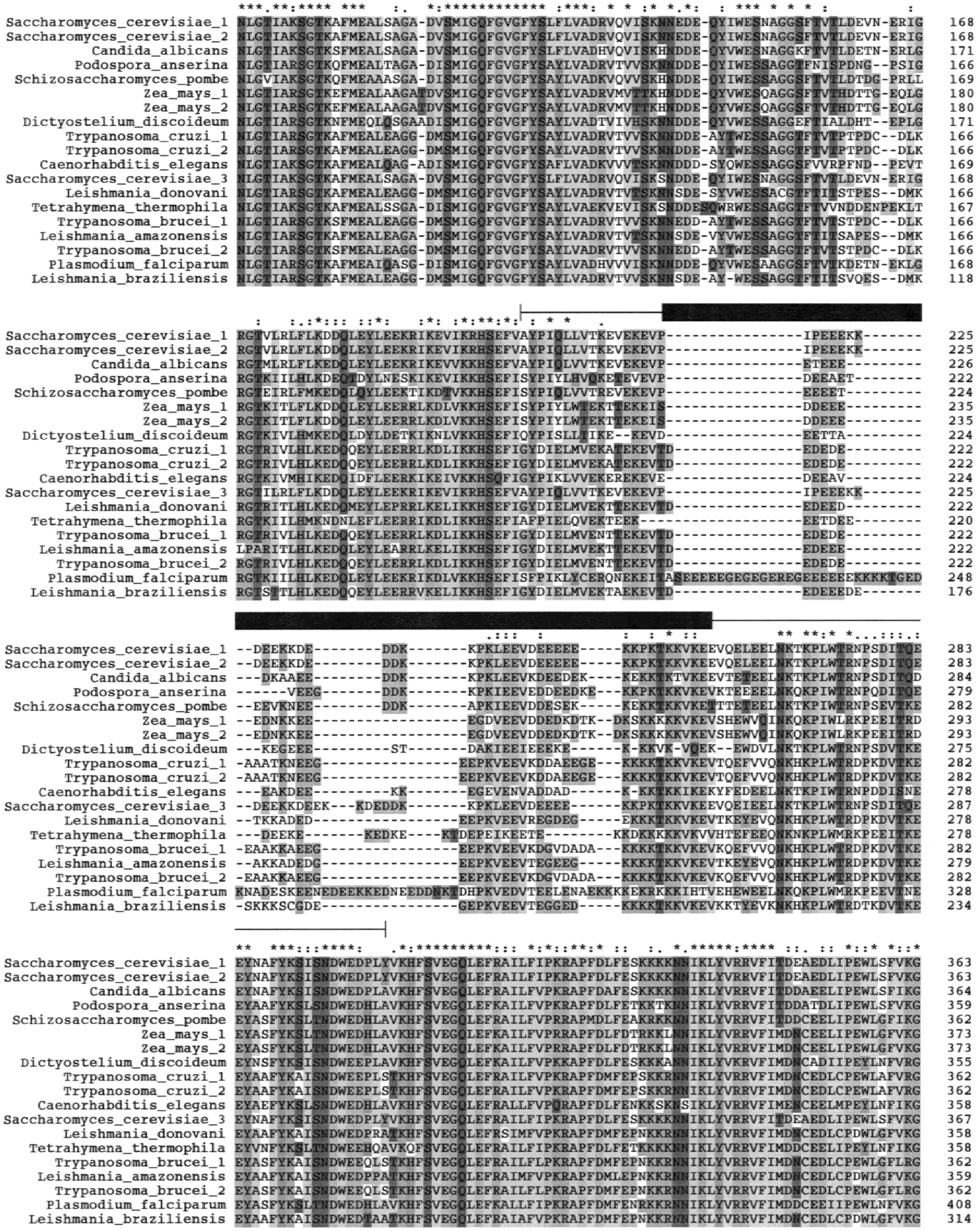

Fig. 1. Alignment of 18 proteins similar to yeast protein hs83. Sequences have been truncated at the amino- and carboxy-termini. The bar shows a repetitive region within the yeast protein and the line shows the corresponding 100-residue segment used in the search (asterisks above the sequence indicate identical amino acids; colons and periods indicate conserved residues; shading is as encoded by Clustal X).

to 160 (the remaining sequences do not have a corresponding peptide region). This similarity breaks down around amino acid 172, where there is a region of small serine and asparagine repeats (results not shown). Thereafter, from amino acid 781 to 1,005, all sequences show great similarity again (Fig. 2). After this, the trk1 sequence has a region with aspartic acid, glutamic acid, and lysine repeats, and once again in this region the similarity of the sequences declines (Fig. 2 shows the alignment surrounding this repeat). From amino acid 1,077 until approaching the carboxy-terminus, these sequences show high similarity. In this case, only the yeast

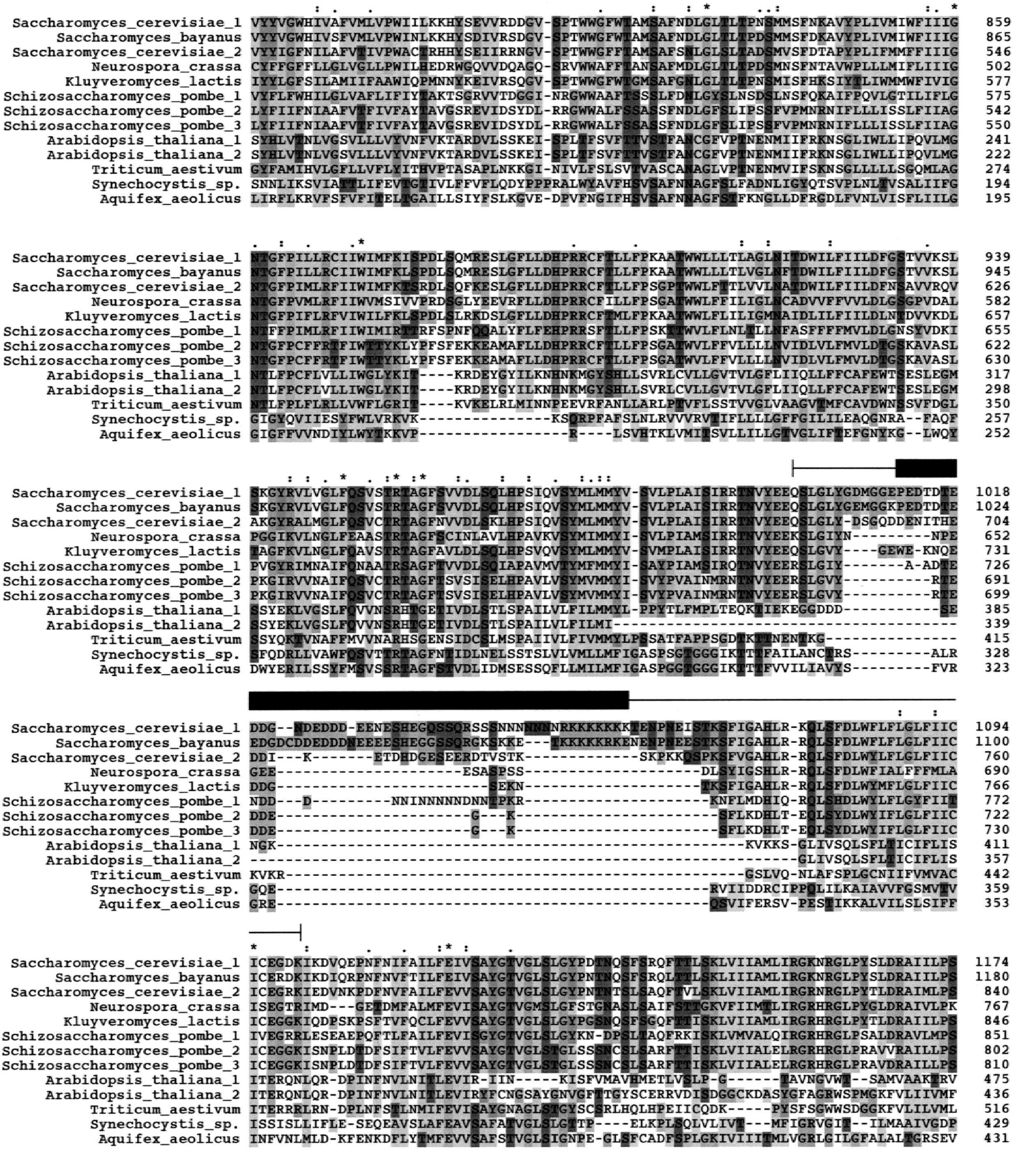
Saccharomyces_cerevisiae_1 VYYVGWHIVAFVMLVPWIILKKHYSEVVRDDGV-SPTWWGFWTAMSAFNDLGLTLTPNSMMSFNKAVYPLIVMIWFIIIG 859
Saccharomyces_bayanus VYYVGWHIVSFVMLVPWINLKKHYSDIVRSDGV-SPTWWGFWTAMSAFNDLGLTLTPDSMMSFDKAVYPLIVMIWFIIIG 865
Saccharomyces_cerevisiae_2 VYYIGFNILAFVTIVPWACTRHHYSEIIRRNGV-SPTWWGFFTAMSAFSNLGLSLTADSMVSFDTAPYPLIFMMFFIIIG 546
Neurospora_crassa CYFFGFFLLGLVGLLPWILHEDRWGQVVDQAGQ-SRVWWAFFTANSAFMDLGLTLTPDSMNSFNTAVWPLLLMIFLIIIG 502
Kluyveromyces_lactis IYYLGFSILAMIIFAAWIQPMNNYKEIVRSQGV-SPTWWGFWTGMSAFGNLGLTLTPNSMISFHKSIYTLIWMMWFIVIG 577
Schizosaccharomyces_pombe_1 VYFLFWHILGLVAFLIFIYTAKTSGRVVTDGGI-NRGWWAAFTSSSLFDNLGYSLNSDSLNSFQKAIFPQVLGTILIFLG 575
Schizosaccharomyces_pombe_2 LYFIIFNIAAFVTFIVFAYTAVGSREVIDSYDL-RRGWWALFSSASSFNDLGFSLIPSSFVPMNRNIFLLLISSLFIIAG 542
Schizosaccharomyces_pombe_3 LYFIIFNIAAFVTFIVFAYTAVGSREVIDSYDL-RRGWWALFSSASSFNDLGFSLIPSSFVPMNRNIFLLLISSLFIIAG 550
Arabidopsis_thaliana_1 SYHLVTNLVGSVLLLVYVNFVKTARDVLSSKEI-SPLTFSVFTTVSTFANCGFVPTNENMIIFRKNSGLIWLLIPQVLMG 241
Arabidopsis_thaliana_2 SYHLVTNLVGSVLLLVYVNFVKTARDVLSSKEI-SPLTFSVFTTVSTFANCGFVPTNENMIIFRKNSGLIWLLIPQVLMG 222
Triticum_aestivum GYFAMIHVLGFLLVFLYITHVPTASAPLNKKGI-NIVLFSLSVTVASCANAGLVPTNENMVIFSKNSGLLLLLSGQMLAG 274
Synechocystis_sp. SNNLIKSVIATTLIFEVTGTIVLFFVFLQDYPPPRALWYAVFHSVSAFNNAGFSLFADNLIGYQTSVPLNLTVSALIIFG 194
Aquifex_aeolicus LIRFLKRVFSFVFITELTGAILLSIYFSLKGVE-DPVFNGIFHSVSAFNNAGFSTFKNGLLDFRGDLFVNLVISFLIILG 195
Saccharomyces_cerevisiae_1 NTGFPILLRCIIWIMFKISPDLSQMRESLGFLLDHPRRCFTLLFPKAATWWLLLTLAGLNITDWILFIILDFGSTVVKSL 939
Saccharomyces_bayanus NTGFPILLRCIIWIMFKLSPDLSQMRESLGFLLDHPRRCFTLLFPKAATWWLLLTLVGLNFTDWILFIILDFGSTVVKSL 945
Saccharomyces_cerevisiae_2 NTGFPIMLRFIIWIMFKTSRDLSQFKESLGFLLDHPRRCFTLLFPSGPTWWLFTTLVVLNATDWILFIILDFNSAVVRQV 626
Neurospora_crassa NTGFPVMLRFIIWVMSIVVPRDSGLYEEVRFLLDHPRRCFILLFPSGATWWLFFILIGLNCADVVFFVVLDLGSGPVDAL 582
Kluyveromyces_lactis NTGFPIFLRFVIWILFKLSPDLSLRKDSLGFLLDHPRRCFTMLFPKAATWWLFLILIGMNAIDLILFIILDLNTDVVKDL 657
Schizosaccharomyces_pombe_1 NTFFPIMLRFIIWIMIRTTRFSPNFQQALYFLFEHPRRSFTLLFPSKTTWVLFLNLTLLNFASFFFFMVLDLGNSYVDKI 655
Schizosaccharomyces_pombe_2 NTGFPCFFRTFIWTTYKLYPFSFEKKEAMAFLLDHPRRCFTLLFPSGATWVLFFVLLLLNVIDLVLFMVLDTGSKAVASL 622
Schizosaccharomyces_pombe_3 NTGFPCFFRTFIWTTYKLYPFSFEKKEAMAFLLDHPRRCFTLLFPSGATWVLFFVLLLLNVIDLVLFMVLDTGSKAVASL 630
Arabidopsis_thaliana_1 NTLFPCFLVLLIWGLYKIT----KRDEYGYILKNHNKMGYSHLLSVRLCVLLGVTVLGFLIIQLLFFCAFEWTSESLEGM 317
Arabidopsis_thaliana_2 NTLFPCFLVLLIWGLYKIT----KRDEYGYILKNHNKMGYSHLLSVRLCVLLGVTVLGFLIIQLLFFCAFEWTSESLEGM 298
Triticum_aestivum NTLFPLFLRLLVWFLGRIT----KVKELRLMINNPEEVRFANLLARLPTVFLSSTVVGLVAAGVTMFCAVDWNSSVFDGL 350
Synechocystis_sp. GIGYQVIIESYFWLVRKVK---------------KSQRPFAFSLNLRVVVRVTIFLLLLGFFGILILEAQGNRA--FAQF 257
Aquifex_aeolicus GIGFFVVNDIYLWYTKKVP-----------------R----LSVHTKLVMITSVLLILLGTVGLIFTEFGNYKG--LWQY 252
Saccharomyces_cerevisiae_1 SKGYRVLVGLFQSVSTRTAGFSVVDLSQLHPSIQVSYMLMMYV-SVLPLAISIRRTNVYEEQSLGLYGDMGGEPEDTDTE 1018
Saccharomyces_bayanus SKGYRVLVGLFQSVSTRTAGFSVVDLSQLHPSIQVSYMLMMYV-SVLPLAISIRRTNVYEEQSLGLYGEMGGKPEDTDTE 1024
Saccharomyces_cerevisiae_2 AKGYRALMGLFQSVCTRTAGFNVVDLSKLHPSIQVSYMLMMYV-SVLPLAISIRRTNVYEEQSLGLY-DSGQDDENITHE 704
Neurospora_crassa PGGIKVLNGLFEAASTRTAGFSCINLAVLHPAVKVSYMIMMYI-SVLPIAMSIRRTNVYEEKSLGIYN---------NPE 652
Kluyveromyces_lactis TAGFKVLNGLFQAVSTRTAGFAVLDLSQLHPSVQVSYMLMMYV-SVMPLAISIRRTNVYEEQSLGVY----GEWE-KNQE 731
Schizosaccharomyces_pombe_1 PVGYRIMNAIFQNAATRSAGFTVVDLSQIAPAVMVTYMFMMYI-SAYPIAMSIRQTNVYEERSLGIY-------A-ADTE 726
Schizosaccharomyces_pombe_2 PKGIRVVNAIFQSVCTRTAGFTSVSISELHPAVLVSYMVMMYI-SVYPVAINMRNTNVYEERSLGVY----------RTE 691
Schizosaccharomyces_pombe_3 PKGIRVVNAIFQSVCTRTAGFTSVSISELHPAVLVSYMVMMYI-SVYPVAINMRNTNVYEERSLGVY----------RTE 699
Arabidopsis_thaliana_1 SSYEKLVGSLFQVVNSRHTGETIVDLSTLSPAILVLFILMMYL-PPYTLFMPLTEQKTIEKEGGDDD-----------SE 385
Arabidopsis_thaliana_2 SSYEKLVGSLFQVVNSRHTGETIVDLSTLSPAILVLFILMI------------------------------------- 339
Triticum_aestivum SSYQKTVNAFFMVVNARHSGENSIDCSLMSPAIIVLFIVMMYLPSSATFAPPSGDTKTTNENTKG--------------- 415
Synechocystis_sp. SFQDRLLVAWFQSVTTRTAGFNTIDLNELSSTSLVLVMLLMFIGASPSGTGGGIKTTTFAILANCTRS---------ALR 328
Aquifex_aeolicus DWYERILSSYFMSVSSRTAGFSTVDLIDMSESSQFLLMILMFIGASPGGTGGGIKTTTFVVILIAVYS---------FVR 323
Saccharomyces_cerevisiae_1 DDG--NDEDDD-EENESHEGQSSQRSSSNNNNNNNRKKKKKKKTENPNEISTKSFIGAHLR-KQLSFDLWFLFLGLFIIC 1094
Saccharomyces_bayanus EDGDCDDEDDDNEEEESHEGGSSQRGKSKKE---TKKKKKRKENENPNEESTKSFIGAHLR-RQLSFDLWFLFLGLFIIC 1100
Saccharomyces_cerevisiae_2 DDI---K-------ETDHDGESEERDTVSTK-------------SKPKKQSPKSFVGAHLR-RQLSFDLWYLFLGLFIIC 760
Neurospora_crassa GEE---------------------ESASPSS------------------DLSYIGSHLR-RQLSFDLWFIALFFFMLA 690
Kluyveromyces_lactis DDG-------------------------SEKN------------------TKSFIGAHLR-RQLSFDLWYMFLGLFIIC 766
Schizosaccharomyces_pombe_1 NDD---D---------NNINNNNNDNNTPKR-------------------KNFLMDHIQ-RQLSHDLWYLFLGYFIIT 772
Schizosaccharomyces_pombe_2 DDE----------------------G---K----------------------SFLKDHLT-EQLSYDLWYIFLGLFIIC 722
Schizosaccharomyces_pombe_3 DDE----------------------G---K----------------------SFLKDHLT-EQLSYDLWYIFLGLFIIC 730
Arabidopsis_thaliana_1 NGK---KVKKS-GLIVSQLSFLTICIFLIS 411
Arabidopsis_thaliana_2 --GLIVSQLSFLTICIFLIS 357
Triticum_aestivum KVKR---GSLVQ-NLAFSPLGCNIIFVMVAC 442
Synechocystis_sp. GQE--RVIIDDRCIPPQLILKAIAVVFGSMVTV 359
Aquifex_aeolicus GRE--QSVIFERSV-PESTIKKALVILSLSIFF 353
Saccharomyces_cerevisiae_1 ICEGDKIKDVQEPNFNIFAILFEIVSAYGTVGLSLGYPDTNQSFSRQFTTLSKLVIIAMLIRGKNRGLPYSLDRAIILPS 1174
Saccharomyces_bayanus ICERDKIKDIQRPNFNVFTILFEIVSAYGTVGLSLGYPNTNQSFSRQLTTLSKLIIIAMLIRGKNRGLPYSLDRAIILPS 1180
Saccharomyces_cerevisiae_2 ICEGRKIEDVNKPDFNVFAILFEVVSAYGTVGLSLGYPNTNTSLSAQFTVLSKLVIIAMLIRGRNRGLPYTLDRAIMLPS 840
Neurospora_crassa ISEGTRIMD---GETDMFALMFEVISAYGTVGMSLGFSTGNASLSAIFSTTGKVFIIMTLIRGRHRGLPYGLDRAIVLPK 767
Kluyveromyces_lactis ICEGGKIQDPSKPSFTVFQCLFEVVSAYGTVGLSLGYPGSNQSFSGQFTTISKLVIIAMLIRGRHRGLPYTLDRAIILPS 846
Schizosaccharomyces_pombe_1 IVEGRRLESEAEPQFTLFAILFEVISGYGTVGLSLGYKN-DPSLTAQFRKISKLVMVALQIRGRHRGLPSALDRAVLMPS 851
Schizosaccharomyces_pombe_2 ICEGGKISNPLDTDFSIFTVLFEVVSAYGTVGLSTGLSSSNCSLSARFTTISKLVIIALELRGRHRGLPRAVVRAILLPS 802
Schizosaccharomyces_pombe_3 ICEGGKISNPLDTDFSIFTVLFEVVSAYGTVGLSTGLSSSNCSLSARFTTISKLVIIALELRGRHRGLPRAVDRAILLPS 810
Arabidopsis_thaliana_1 ITERQNLQR-DPINFNVLNITLEVIR-IIN-----KISFVMAVHMETLVSLP-G------TAVNGVWT--SAMVAAKTRV 475
Arabidopsis_thaliana_2 ITERQNLQR-DPINFNVLNITLEVIRYFCNGSAYGNVGFTTGYSCERRVDISDGGCKDASYGFAGRWSPMGKFVLIIVMF 436
Triticum_aestivum ITERRRLRN-DPLNFSTLNMIFEVISAYGNAGLSTGYSCSRLHQLHPEIICQDK-----PYSFSGWWSDGGKFVLILVML 516
Synechocystis_sp. ISSISLLIFLE-SEQEAVSLAFEAVSAFATVGLSLGTTP----ELKPLSQLVLIVT---MFIGRVGIT--ILMAAIVGDP 429
Aquifex_aeolicus INFVNLMLD-KFENKDFLYTMFEVVSAFSTVGLSIGNPE-GLSFCADFSPLGKIVIIITMLVGRLGILGFALALTGRSEV 431

Fig. 2. Alignment of 12 proteins similar to yeast protein trk1 (see Fig. 1). Sequences have been truncated at the amino- and carboxy-termini.

sequences contain the repetitive region. Other species lack it altogether, even in proteins with high levels of similarity.

These and other examples indicate that the repeats are often present at or near the beginning or the end of a protein and that they tend to change relatively quickly in comparison to the remainder of the protein. When located in the middle of the protein, the unusual repeats are often surrounded by other regions of the protein that change at a slower rate.

To make these observations more quantitative we compared the proteins from yeast and aligned them with their closest nonyeast homologue (as judged by a BLAST expect value criterion). The differences in the percentage of amino acid replacements within the 100-residue segment versus outside of this segment were compared. The differences were also compared in the percentage of indels and for the percentage of replacements plus indels within versus outside the 100-residue segment.

Table 4. Relative evolutionary rates

Number of genes	Slope	t	$P < t$	r^2
Amino acid replacements				
3,014	0.0042 ± 0.0008	5.206	0.000	0.0089
Insertions/deletions				
3,230	0.0023 ± 0.0013	6.686	0.000	0.0137
Both				
3,230	0.0141 ± 0.0021	6.842	0.000	0.0143

A regression of this difference versus the number of times the hundred-residue segment has similar segments in other proteins has a significantly positive slope in all three cases $Pr < 0.0005$ (Table 4; based on 3,014 observations for amino acid replacements, 3,230 observations for indels, and 3,230 observations for both—the number of observations for amino acid replacements is less than the others because for many of the proteins the 100-residue segment had no complimentary sequence in the closest homologue). The positive slope indicates that the more frequent the segment is represented in the genome, the more likely that it will have a larger evolutionary rate than the surrounding protein. However, the coefficient of determination, r^2, in each case is very small (0.0089, 0.0137, 0.0143, respectively), indicating that there are many other factors that influence their relative evolutionary rate and that although the slope is significantly positive the frequency of the segment has little power to explain all of the variation observed in the evolutionary rate.

Discussion

When DNA reassociation experiments were first done it was discovered that most eukaryotes have large amounts of repetitive DNA (Britten and Kohne 1968). Repetitive DNA sequences have now been discovered in all free living organisms. Even the small bacterial genome of *Mycoplasma genitalium* has DNA repeats (Hancock 1996). It is less commonly observed that protein sequences are also repetitive in nature. But simple sequence repeats are in fact the most commonly shared pattern between all of the genomic proteins of yeast. It has been shown that 14% of all yeast proteins show significant similarity to a poly-S segment of protein and a total of 21% of all yeast proteins have a segment that has significant similarity to either poly-S, poly-E, poly-D, poly-Q, or poly-N (Golding 1999). Not all proteins that contain repetitive simple sequence are immediately apparent to the eye because via a PAM matrix functionally equivalent amino acids will score nearly as high as the identical amino acid.

The origin of these simple sequence repeats are of interest. The eukaryotes are thought to have evolved from a common ancestor with archaebacteria (Woese et al. 1990). Therefore we have analyzed four different prokaryotic genomes to determine if shared repeats are common among their proteins. These genomes were chosen from taxonomic groups as widely scattered as possible: representatives from the Gram-negative bacteria, from the Gram-positive bacteria, and from the archaebacteria.

For this analysis we did not wish to include closely related proteins. One of the most commonly accepted origins for new proteins is gene duplication from a previously existing protein. This process will lead to large multigene families when duplication occurs rapidly. Over evolutionary time these duplicate proteins will diverge in sequence and can, given the correct replacements, encode a different function. For example, many of the proteins in the yeast genome are known to be ancient duplications (Wolfe and Shields 1997). Therefore, all proteins were pairwise aligned to eliminate proteins with similar segments due only to recent shared ancestry. Any entries with more than 20% identity throughout their entire length were eliminated. Because the proteins (including translated open reading frames) from each organism were analyzed separately, the comparisons presented here are strictly intraspecific.

The results presented here suggest that these repeats are a uniquely eukaryotic feature of protein structure. This is different from the simple presence of low-complexity sequence. Low-complexity sequence is present in the proteins of both prokaryotes and eukaryotes (Marcotte et al. 1998). But although it is present in prokaryotes, it does not seem to contribute in any substantial way to the major shared sequences between distinct proteins. Repeats within DNA sequences are more common in larger organisms and there is a good relationship between DNA repetitiveness and genome length (Hancock 1995). But this is not likely to be the reason that repetitive simple sequences are missing in the prokaryotic protein sequences. Indeed there is no a priori reason to expect that a similar relationship holds for protein repeats. Although yeast is a single-celled organism with a small genome, it has more of this type of simple protein sequence than does the multicellular nematode with a larger genome.

Many simple sequence protein repeats have been observed previously. The *opa* repeats originally discovered in insects are the most famous of these repeats. *Opa* repeats are simple sequence repeats consisting of poly-Q with, for example, 31 tandem residues present in an *opa* repeat in the notch locus. *Opa*-like repeats have been discovered in *Drosophila* (Wharton et al. 1985), medflies (Siden-Kiamos et al. 1993), and mice (Duboule et al. 1987; Persengiev and Kilpatrick 1997). They have been suggested to be characteristic of developmentally regulated genes (Wharton et al. 1985). Other simple sequence repeats are the alanine-rich antifreeze proteins of fish (Lin and Gross 1981), alanine tracts in molluscan shell framework proteins (Sudo et al. 1997), or the polyglutamine repeats in murine GRP-1 (Cox et al. 1996). Many of these tandem repeats have been noticed in individual proteins and often, their presence has been noted as rather unusual (O'Hara et al. 1988; Vai et al. 1991; White et al. 1991; Heinonen and Pearlman 1994; Wootton 1994; Di Como et al. 1995; Yamamoto et al. 1995; Cox et al. 1996; Sudo et al. 1997). But the length and high repetitive frequency as illustrated here is not commonly appreciated.

The high frequency of these repeats in diverse proteins suggests that they must either have an important, broadly based function or that they are simply dispensible for the protein and happen to be residues that will not disrupt the remainder of the protein. There are five features of these sequences that suggest that to some degree the latter may be true. The first of these is that the nature of the repetitive amino acids seems to argue against a specific uniform function for all repeats. There are no unusual characteristics of these particular amino acids. They are not particularly large or small amino acids. Nor are they particularly unreactive. Both D and E are acidic residues, S (and T) are hydroxyl residues, and N and Q are amide residues. The suggestion has been made that simple sequence in proteins favors hydrophilic amino acid residues (Marcotte et al. 1998) but some repetitive simple sequences are known to be composed of the aliphatic residues alanine and leucine. It is difficult to see how these repeats would form useful secondary or tertiary structures. The disordered tertiary structures that they might form (Newfeld et al. 1994) could favor these particular amino acids residues. Unfortunately, the tertiary structure of these elements has received comparatively little study but glutamine homopolymers are known to form stable β-sheets (Perutz et al. 1994). If the composition of these tandem repeats impart a distinct function that would explain their high frequency, it is not readily apparent.

Second, the rate of evolution within the repeats is higher than the rate of evolution outside of the repeats and the rate is positively correlated with the frequency of the repeat throughout the genome. Newfeld et al. (1993, 1991) have demonstrated this difference in evolutionary rate within versus outside repetitive simple sequence regions of the *Drosophila mastermind* gene. The results presented here show that their conclusion holds on a genome-wide level. The most likely cause of this difference in evolutionary rate is a significant difference in functional constraint. Not unexpectedly, many additional but currently unknown factors also strongly influence these rates. The generation of slippage mutations in these repetitive simple sequences are no doubt significant contributors to the higher evolutionary rate (Newfeld et al. 1994).

Third, if these repeats are uniformly functional in all eukaryotes, it might be unexpected to have large frequency differences between species. Yet from the partial results for *D. discoideum*, more than 32% of all its proteins contain segments with significant similarity to a single repeat, and *A. thaliana* has only 5.4% of its proteins similar to one repeat. Fourth, the substance of the repeat changes between eukaryotic species. *C. elegans* had threonine, proline, and glycine repeats that are not observed as common repeats in yeast. It is difficult to see how the repeats could encode specific functions if the identity of the amino acid residue was not important. Nor why one organism would require a very common feature in its genome that is not required by another.

Last, the repeat region of some proteins might be deleted without affecting the function of the protein. Gatti et al. (1994) both deleted and duplicated a 36-amino acid serine-rich region from yeast protein gp115. Neither construction outwardly affected the function of the protein despite the serines being targets of O-glycosylation. Similarly, Sumiyama et al. (1996) found alanine, glycine, and proline repeats in the mammalian Brain-1 and Brain-2 class III POU transcription factor genes. Yet their nonmammalian homologues lack these repeats altogether.

However, it is also apparent that at least some of these repetitive simple sequences have been assigned critical functions. It has been suggested that some of these regions have important interactions or functions (Wootton 1994) in a variety of roles. Because some of these regions appear to be functionally necessary even though their primary sequence may be highly variable it would suggest that in these cases, it may be that the presence of a repetitive simple sequence is more important than its primary sequence composition. In other cases, it is possible that the function of these repeats is to simply serve as spacers between other protein motifs—the protein equivalent of junk DNA—but this cannot yet be confirmed. In either case, the repeats give much greater flexibility in protein structure and greater variability between species in protein sequence than would otherwise be possible.

Acknowledgments. This work was supported by a Natural Sciences and Engineering Research Council of Canada grant to GBG. We thank Dr. R. Morton and two anonymous reviewers for their comments on previous drafts. GBG also thanks Dr. P. Green for providing the reference to Wootton (1994).

References

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

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Borkovich KA, Farrelly FW, Finkelstein DB, Taulien J, Lindquist S (1989) hsp82 is an essential protein that is required in higher concentrations for growth of cells at higher temperatures. Mol Cell Biol 9:3919–3930

Britten RJ, Kohne DE (1968) Repeated sequences in DNA. Science 161:529–540

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage

AR, Dougherty BA, Tomb J, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JD, Geoghagen NS, Weidman JF, Fuhrmann JL, Nguyen DT, Utterback T, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Hurst MA, Roberts KM, Kaine BB, Borodovsky M, Klenk HP, Fraser CM, Smith HO, Woese CR, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science 273:1058–1073

Cox GW, Taylor LS, Willis JD, Melillo G, White R, Anderson SK, Lin JJ (1996) Molecular cloning and characterization of a novel mouse macrophage gene that encodes a nuclear protein comprising polyglutamine repeats and interspersing histidines. J Biol Chem 271: 25515–25523

Creighton T (1993) Proteins: structures and molecular properties, 2nd ed. W.H. Freeman and Company, New York, NY

de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W (1996) Intron positions correlate with module boundaries in ancient proteins. Proc Natl Acad Sci USA 93:14632–14636

Di Como CJ, Bose R, Arndt KT (1995) Overexpression of SIS2, which contains an extremely acidic region, increases the expression of SWI4, CLN1 and CLN2 in sit4 mutants. Genetics 139:95–107

Doolittle RF (1995) The multiplicity of domains in proteins. Ann Rev Biochem 64:287–314

Dorit RL, Schoenbach L, Gilbert W (1990) How big is the universe of exons? Science 250:1377–1382

Duboule D, Haenlin M, Galliot B, Mohier E (1987) DNA sequences homologous to the *Drosophila* opa repeat are present in murine mRNAs that are differentially expressed in fetuses and adult tissues. Mol Cell Biol 7:2003–2006

Gaber RF, Styles CA, Fink GR (1988) TRK1 encodes a plasma membrane protein required for high-affinity potassium transport in *Saccharomyces cerevisiae.* Mol Cell Biol 8:2848–2859

Gatti E, Popolo L, Vai M, Rota N, Alberghina L (1994) O-linked oligosaccharides in yeast glycosyl phosphatidylinositol-anchored protein gp115 are clustered in a serine-rich region not essential for its function. J Biol Chem 269:19695–19700

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci USA 94:7698–7703

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546

Goffeau A, Aert R, Agostini-Carbone M, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M, Alexandraki D, Aljinovic G, Allen E, et al. (1997) The yeast genome directory. Nature 387:1–105

Golding GB (1999) Simple sequence is abundant in eukaryotic proteins. Protein Sci 8:1358–1361

Hancock JM (1995) The contribution of slippage-like processes to genome evolution. J Mol Evol 41:1038–1047

Hancock JM (1996) Simple sequences in a “minimal” genome. Nat Genet 14:14–15

Heinonen TY, Pearlman RE (1994) A germ line-specific sequence element in an intron in *Tetrahymena thermophila.* J Biol Chem 269:17428–17433

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, et al. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis.* Nature 390:249–256

Lin Y, Gross JK (1981) Molecular cloning and characterization of winter flounder antifreeze cDNA. Proc Natl Acad Sci USA 78: 2825–2829

Marcotte E, Pellegrini M, Yeates T, Eisenberg D (1998) A census of protein repeats. J Mol Biol 293:151–160

Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A (1997) Overview of the yeast genome. Nature 387:7–65

Newfeld S, Smoller D, Yedvobnick B (1991) Interspecific comparison of the unusually repetitive *Drosophila* locus *mastermind*. J Mol Evol 32:415–420

Newfeld S, Schmid A, Yedvobnick B (1993) Homopolymer length variation in the *Drosophila* gene *mastermind.* J Mol Evol 37:483–495

Newfeld S, Tachida H, Yedvobnick B (1994) Drive-selection equilibrium: homopolymer evolution in the *Drosophila* gene *mastermind*. J Mol Evol 38:637–641

O’Hara PJ, Horowitz H, Eichinger G, Young ET (1988) The yeast ADR6 gene encodes homopolymeric amino acid sequences and a potential metal-binding domain. Nucleic Acids Res 16:10153–10169

Ohno S (1987) Early genes that were oligomeric repeats generated a number of divergent domains on their own. Proc Natl Acad Sci USA 84:6486–6490

Persengiev SP, Kilpatrick DL (1997) Characterization of a cDNA containing trinucleotide repeat sequences that is highly enriched in spermatogenic cells. Mol Reprod Dev 46:476–481

Perutz M, Johnson T, Suzuki M, Finch J (1994) Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci USA 91:5355–5358

Shaw DR, Richter H, Giorda R, Ohmachi T, Ennis HL (1989) Nucleotide sequences of *Dictyostelium discoideum* developmentally regulated cDNAs rich in (AAC) imply proteins that contain clusters of asparagine, glutamine, or threonine. Mol Gen Genet 218:453–459

Siden-Kiamos I, Favia G, Artiaco D, Saccone G, Furia M, Polito LC, Louis C (1993) Opa-like repeats in the genome of the medfly *Ceratatis capitata.* Genetica 92:43–53

Sudo S, Fujikawa T, Nagakura T, Ohkubo T, Sakaguchi K, Tanaka M, Nakashima K, Takahashi T (1997) Structures of mollusc shell framework proteins. Nature 387:563–564

Sumiyama K, Washio-Watanabe K, Saitou N, Hayakawa T, Ueda S (1996) Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. J Mol Evol 43:170–178

Tanaka T, Kawarabayasi Y, Kikuchi H (1998) Direct submission. DB-JEMBLGenBank databases

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans:* a platform for investigating biology. Science 282:2012–2018

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Vai M, Gatti E, Lacana E, Popolo L, Alberghina L (1991) Isolation and deduced amino acid sequence of the gene encoding gp115, a yeast glycophospholipid-anchored protein containing a serine-rich region. J Biol Chem 266:12242–12248

Wharton KA, Yedvobnick B, Finnerty VG, Artavanis-Tsakonas S (1985) Opa: a novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in *D. melanogaster*. Cell 40:55–62

White MJ, Hirsch JP, Henry SA (1991) The OPI1 gene of *Saccharomyces cerevisiae,* a negative regulator of phospholipid biosynthesis, encodes a protein containing polyglutamine tracts and a leucine zipper. J Biol Chem 266:863–872

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

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

Wooton J (1994) Sequences with “unusual” amino acid compositions. Curr Opin Struct Biol 4:413–421

Yamamoto A, DeWald DB, Boronenkov IV, Anderson RA, Emr SD, Koshland D (1995) Novel PI(4)P 5-kinase homologue, Fab1p, essential for normal vacuole function and morphology in yeast. Mol Biol Cell 6:525–539