JMB

A Census of Protein Repeats

Edward M. Marcotte, Matteo Pellegrini, Todd O. Yeates and David Eisenberg*

Molecular Biology Institute
UCLA-DOE Lab of Structural Biology and Molecular Medicine, University of California, Los Angeles
P.O. Box 951570, Los Angeles
CA, 90095, USA

**Corresponding author*

In this study, we analyzed all known protein sequences for repeating amino acid segments. Although duplicated sequence segments occur in 14% of all proteins, eukaryotic proteins are three times more likely to have internal repeats than prokaryotic proteins. After clustering the repetitive sequence segments into families, we find repeats from eukaryotic proteins have little similarity with prokaryotic repeats, suggesting most repeats arose after the prokaryotic and eukaryotic lineages diverged. Consequently, protein classes with the highest incidence of repetitive sequences perform functions unique to eukaryotes. The frequency distribution of the repeating units shows only weak length dependence, implicating recombination rather than duplex melting or DNA hairpin formation as the limiting mechanism underlying repeat formation. The mechanism favors additional repeats once an initial duplication has been incorporated. Finally, we show that repetitive sequences are favored that contain small and relatively water-soluble residues. We propose that error-prone repeat expansion allows repetitive proteins to evolve more quickly than non-repeat-containing proteins.

Keywords: duplication; protein evolution; genomic analysis; minisatellite; microsatellite

Introduction

Repetitive nucleotide sequences are frequently found in eukaryotic genomes, in loci that have been termed micro- and minisatellites, depending on the repeat length. These regions are often hypermutable, rapidly gaining and losing repeats during the course of evolution (Kruglyak *et al.*, 1998; Buard & Vergnaud, 1997). Although they are usually found in non-coding genomic regions, repeating sequences are also found within genes. Within the protein sequences coded by these genes, the repeats come in considerable variety (reviewed by Heringa, 1998), ranging from repeats of a single amino acid, through three residue short tandem repeats (e.g. in collagen), to the repetition of homologous domains of 100 or more residues (e.g. the domains of antibodies).

Internal repeats have been studied previously in individual proteins (e.g. McLachlan, 1983; Heringa & Argos, 1993), but fast algorithms for surveying all known protein sequences for repeats have only recently been developed (Pellegrini *et al.*, 1999). These algorithms are capable of detecting both tandem, or adjacent repeats, and non-tandem repeats separated by intervening sequence. Protein repeats have implications not only for evolution but also for genome variability (Kachroo *et al.*, 1997) and disease processes (Djian, 1998), such as Huntington's disease. We present here a census of the internal repeats in all known proteins and draw general conclusions about the role of repetition in evolution of proteins.

Results

Proteins having internal repeats

We begin by attempting to measure the abundance of proteins containing internal duplication within available databases. A portion of this analysis entailed the examination of 70,822 proteins of less than 2000 amino acid residues within the SWISS-PROT database (Bairoch & Apweiler, 1998). We tested each sequence for the presence of repeating sequences and found that 14% of the proteins have one or more statistically significant internal duplications, less than half the duplication rate observed in entire genomes (Gerstein, 1997). In

E.M. Marcotte and M. Pellegrini contributed equally to this work.

E-mail address of the corresponding author: david@mbi.ucla.edu addition, the complete genomes of 16 organisms were analyzed, as were subsets of genes in SWISS-PROT from organelle chromosomes and from several eukaryotes.

When we display the fraction of proteins having internal repeats as a function of average protein length, we find strong clustering of sequences from eukaryotes, prokaryotes and archae (Figure 1(a)). On average, eukaryotes have a significantly higher incidence of internal repeats than either prokaryotes or archaea. We also find that proteins from chloroplasts and mitochondria, organelles descended from prokaryotic ancestors, cluster with the prokaryotic proteins. Various categories of eukaryote-specific proteins, such as proteins in the endoplasmic reticulum, golgi, and nucleus, cluster with the eukaryotes.

On average, eukaryotes have longer proteins than prokaryotes and archael organisms (Netzer & Hartl, 1997). We asked whether this length is the primary cause of the clustering found in Figure 1(a). To test this hypothesis, we compared the incidence of internal repeats as a function of sequence length for the three superkingdoms (Figure 1(b)). The results show that the incidence of repeats in each superkingdom is relatively independent of sequence length for sequences up to 500 amino acid residues, suggesting that the different length distributions of eukaryotic and prokaryotic proteins do not account for the clustering in Figure 1(a). Rather, for proteins of the same length, eukaryotic proteins are approximately three times more likely to have internal duplications than prokaryotes, with archaea falling inbetween (see Figure 1(b)). The distribution of archael repeats fits the notion that these organisms have an intermediate evolutionary relationship between prokaryotes

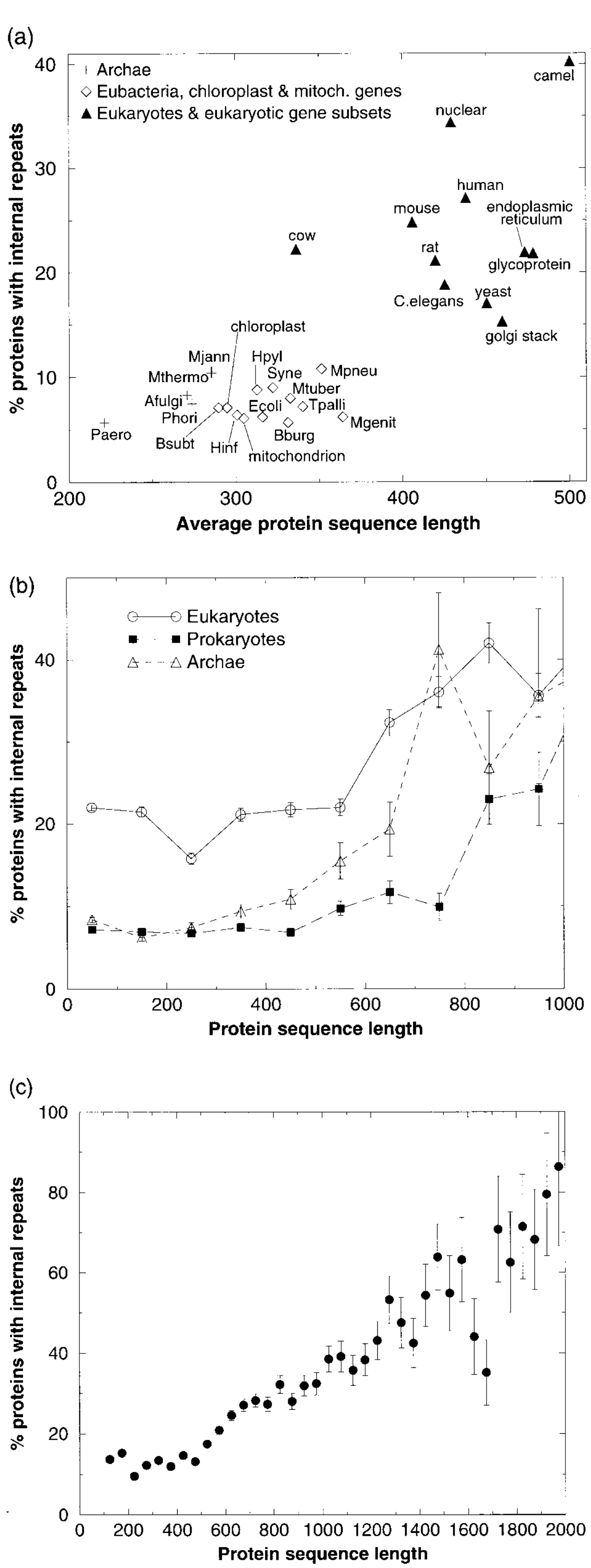


Figure 1. (a) The incidence of internal repeats in proteins from 16 complete genomes and 12 protein subsets from SWISS-PROT 35.0. The fraction of proteins in a genome that contain two or more units of repeating sequence is plotted against the mean protein sequence length in that genome. Eukaryotic protein subsets cluster, as do prokaryotic and archaeal genomes. Proteins with distinctly eukaryotic function (glycoprotein, golgi stack, endoplasmic reticulum) cluster with the eukaryotes; proteins from eukaryotic organelles derived from prokaryotes (chloroplast, mitochondrion) cluster with prokaryotes. (b) Grouping the incidence of repeats in SWISS-PROT 35.0 by superkingdom shows that for proteins of the same length (plotted in 100 residue bins), eukaryotic proteins are three times more likely to have repeats than prokaryotic proteins. In proteins shorter than 500 amino acid residues, the chance of finding repeats shows no dependence on sequence length, although (c) longer proteins, taken from all species and plotted in 50 residue bins, show a roughly linear dependence on length. Errors are modeled as the square-root of the number of proteins in each bin. Afulgi, *Archaeoglobus fulgidus*; Bsubt, *Bacillus subtilis*; Bburg, *Borrelia burgdorferi*; C.elegans, *Caenorhabditis elegans*; Ecoli, *Escherichia coli*; Hinf, *Haemophilus influenzae*; Hpyl, *Helicobacter pylori*; Mthermo, *Methanobacterium thermoautotrophicum*; Mjann, *Methanococcus jannaschii*; Mtuber, *Mycobacterium tuberculosis*; Mgenit, *Mycoplasma genitalium*; Mpneu, *Mycoplasma pneumoniae*; Paero, *Pyrobaculum aerophilum*; Phori, *Pyrococcus horikoshii OT3*; yeast, *Saccharomyces cerevisiae*; Syne, *Synechocystis PCC6803*; Tpalli, *Treponema pallidum*.

and eukaryotes (Koonin *et al.*, 1997). The appearance of repeats in archael proteins does appear to show a dependence upon sequence length. However, it is possible that this is a consequence of the apparent chimeric origins of archae (Koonin *et al.*, 1997), where longer sequences are eukaryotic in origin and the observed length dependence is due to increasing ratios of eukaryotic-like genes.

Unlike shorter proteins, proteins longer than 500 amino acid residues show an incidence of repeats which is correlated with sequence length. In Figure 1(c) we compute the percentage of sequences in SWISS-PROT that contain repeats as a function of their length. Beyond 500 amino acid residues, a linear dependence upon sequence length is seen, suggesting that generation of internal repeats is an important mechanism for producing long proteins.

To ensure that our results are not due to a bias in the algorithm's performance with regard to repetitive sequences from different superkingdoms, the distributions of probability scores are plotted by superkingdom in Figure 2. The repeat-finding algorithm gives similar performance for proteins from each superkingdom with the only significant difference being that for archaea, relatively fewer extremely well-determined repeats are detected (probability of occurring by random chance of $p < 1 \times 10^{-16}$) and more that are near the acceptability threshhold ($p > 1 \times 10^{-3}$). Since we examine average properties of repeats in Figure 2, we cannot rule out a more subtle systematic bias.

Functions of the repeats

Why do eukaryotic genomes code for more proteins with internal repeats than prokaryotic and archael genomes? One possibility is that eukaryotic repeats have functions unique to this superkingdom. To test this hypothesis we identified which classes of proteins have the highest and lowest incidence of repeats. We tabulated the fraction of proteins with a given keyword in the SWISS-PROT database that have one or more internal repeats. The results, shown in Figure 3(a), suggest that the classes of proteins most likely to contain repeats are in fact predominantly unique to eukaryotes. They include connective tissue proteins, cytoskeletal proteins, ribonucleoproteins, muscle proteins, brain and synaptic proteins, and cell adhesion proteins. The classes also include many sets of proteins that share only discrete functional motifs, such as for calcium-binding, in what are otherwise proteins of unrelated sequence and function.

Ancient protein classes that are shared among eukaryotes and prokaryotes appear among the proteins least likely to have repeats. This finding supports the notion that repeats are recent evolutionary events. The list of these proteins is shown in Figure 3(b), and includes proteins from central metabolic pathways, proteins involved in sugar metabolism, DNA synthesis, transport, amino acid biosynthesis, and photosynthesis.

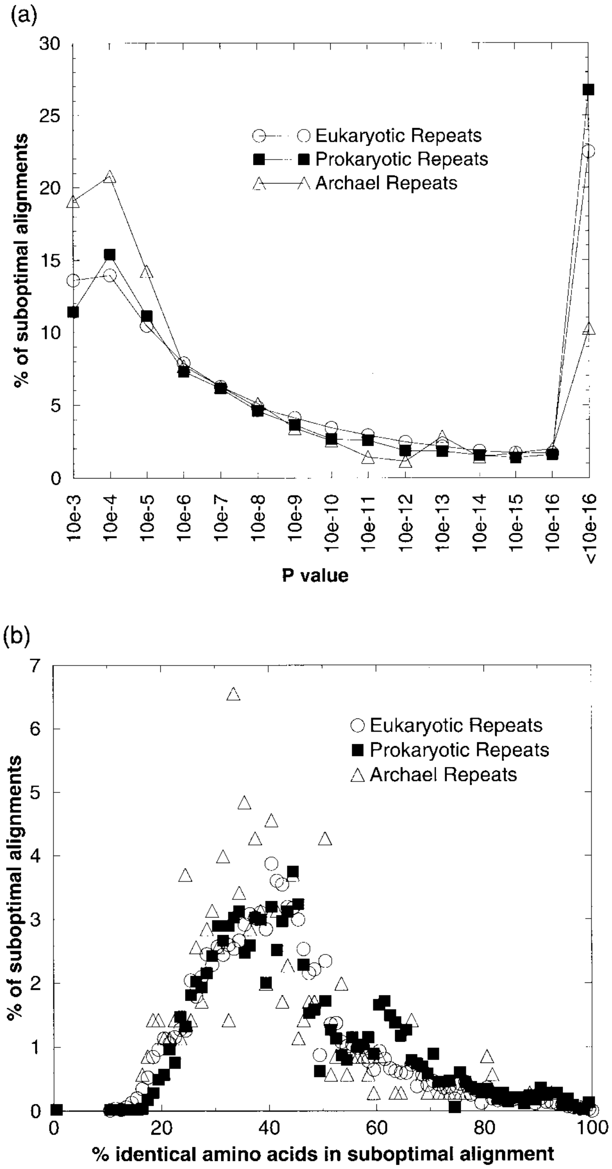


Figure 2. Tests to rule out bias in the algorithm's performance in detecting repetitive sequences in the different superkingdoms. (a) The distribution of probability scores for the statistically significant suboptimal alignments detected in proteins of SWISS-PROT. The *P* values show similar distributions for each superkingdom, although archael proteins have fewer repeats of extremely high significance and proportionally more repeats of lower significance. (b) The distribution of amino acid sequence identity in the statiscally significant sub-optimal alignments detected in proteins of SWISS-PROT. Again, the alignment statistics are similar for each superkingdom, tending to rule out bias.

Of the 4369 protein sequences annotated with the keyword "REPEAT", we identify only 77%, although we find repetitive sequences in 6548 additional proteins that were not marked as repeat-containing. Examining why the automatic repeat-finding algorithm missed proteins marked as containing repeats in SWISS-PROT reveals several

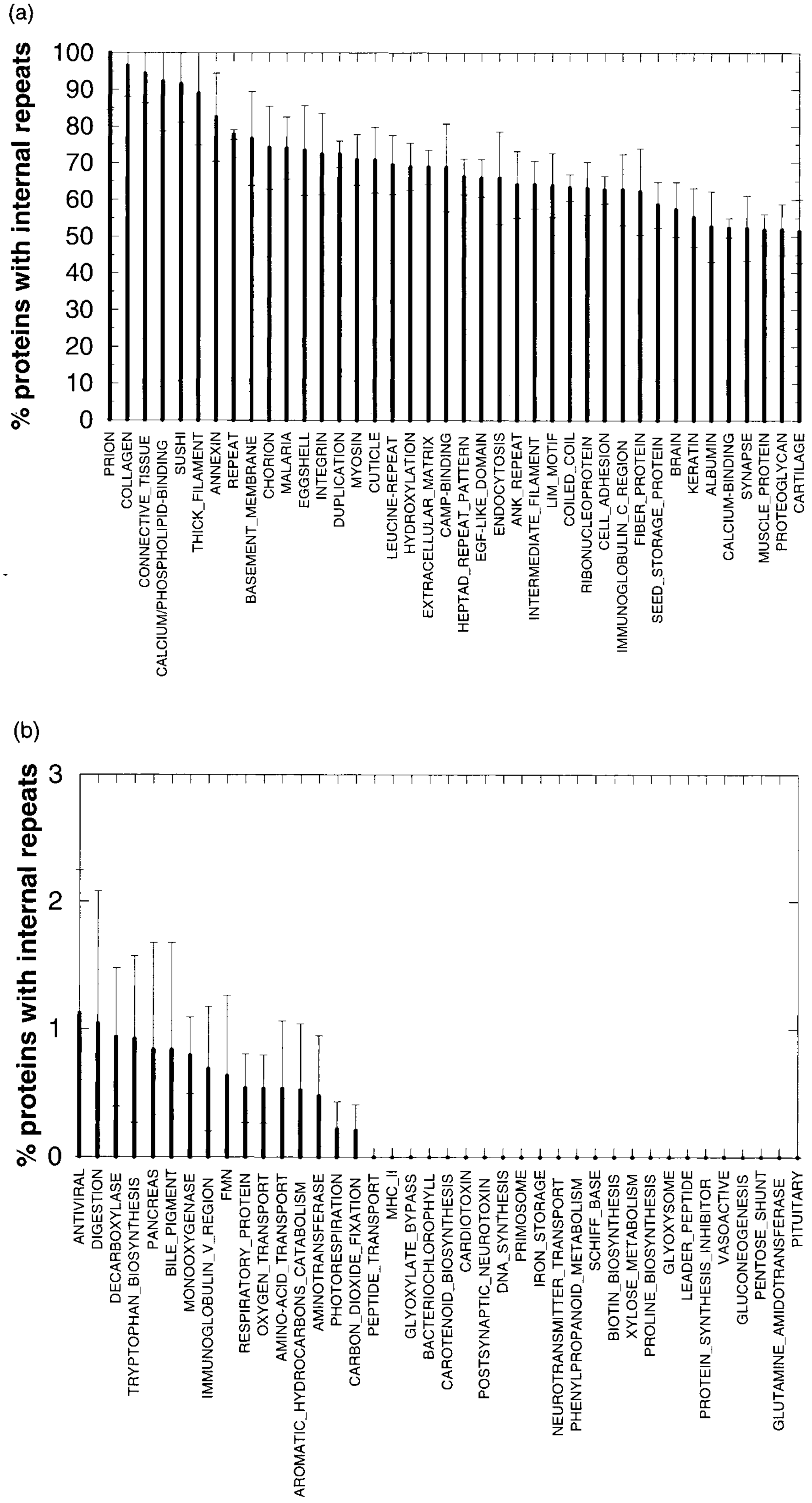


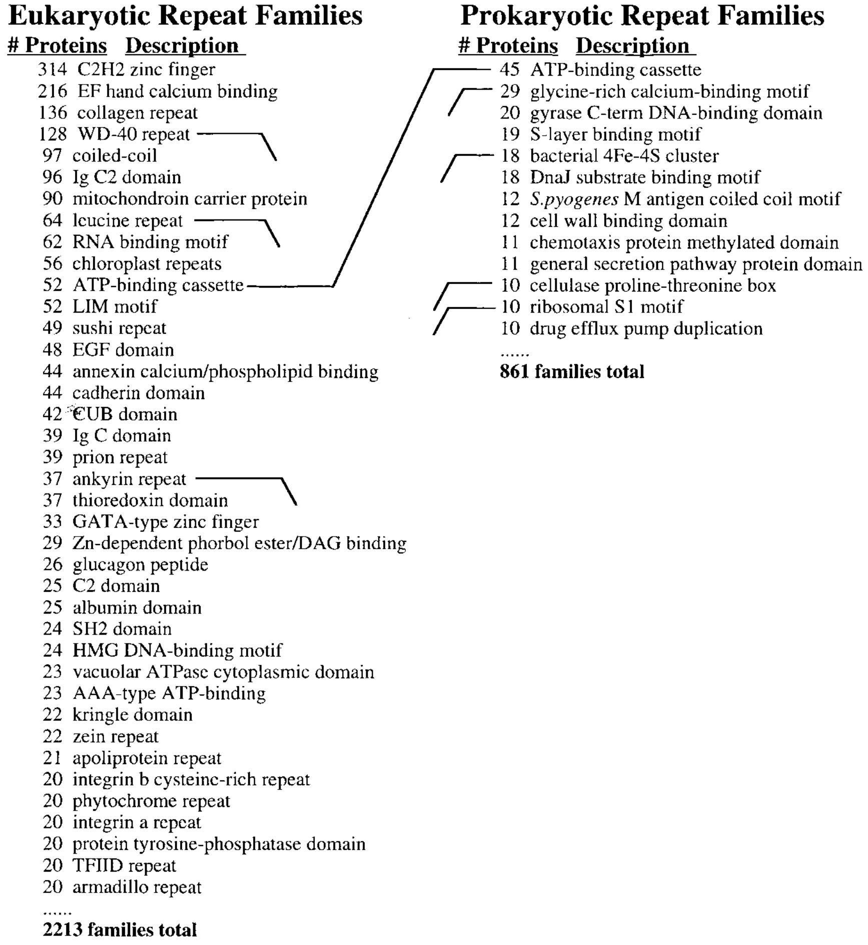
Figure 3. Repeating sequences. Identification of proteins with the highest and lowest incidence of detectable internal repeats. Sets of proteins sharing a common SWISS-PROT keyword (>40 proteins per set) were ranked by the percentage of proteins in the set with at least one internal repeat. (a) The 40 sets with the highest occurence of repeats and (b) the 40 sets with the lowest occurence of repeats out of 460 sets.

trends: (1) some repeated domains have very weak sequence conservation, such as immunoglobulin C2 domains, cystathionine β-synthase domains, and coiled-coil domains. While we detect the majority of these, they are occasionally difficult to detect by sequence alignment methods (Brenner *et al.*, 1995) and are often known because of methods like Pfam (Bateman *et al.*, 1999) or PROSITE (Hofmann *et al.*, 1999) that search explicitly for predetermined domains with domain-specific profiles. (2) Also difficult to detect by sequence alignment methods are domains defined by their pattern of cysteine residues with very little additional sequence conservation, such as tumor-necrosis factor receptor repeats, cystatin-like domains, and LIM motifs. Again, a method based upon domain-specific profiles that explicitly searches for these domains would be expected to perform better than the relatively unbiased search we perform. (3) However, many of the REPEAT-containing sequences missed by our algorithm encode only protein sequence fragments containing a single unit, but which are known to be repeated in other full-length sequence homologs. For example, 91 such fragments occur for the KAZAL-repeat-containing ovomucoid protein family; the fragments each contain only one KAZAL unit but are labeled as containing a repeat.

Differences of repeats in the three superkingdoms

Among the classes of repeat-containing proteins, we examined in more detail whether the repeating fragments themselves were conserved across superkingdoms. To this end, we searched for homologs between the set of eukaryotic and prokaryotic repetitive sequence fragments. We found little sequence similarity between these sets. Only 4% of the eukaryotic repeats have homologs among the prokaryotic repeats. Also, the eukaryotic repeats showed equally little overlap (4%) with non-repeating prokaryotic proteins. The majority of repeats shared by both eukaryotes and prokaryotes were identified as ATP-binding cassettes from ABC-transporter proteins.

To illustrate further the differences between repeat fragments from the different superkingdoms, we clustered eukaryotic and prokaryotic repeats into homologous families (Figure 4). In accordance with our previous results, prokaryotic repeats clustered into 861 families, only about one-third of the 2213 eukaryotic families. Only 14 well-populated prokaryotic repeat families (with ten or more protein sequences) were found, holding 17% of the sequences with repeats. The family with the



Eukaryotic Repeat Families

# Proteins	Description
314	C2H2 zinc finger
216	EF hand calcium binding
136	collagen repeat
128	WD-40 repeat
97	coiled-coil
96	Ig C2 domain
90	mitochondroin carrier protein
64	leucine repeat
62	RNA binding motif
56	chloroplast repeats
52	ATP-binding cassette
52	LIM motif
49	sushi repeat
48	EGF domain
44	annexin calcium/phospholipid binding
44	cadherin domain
42	CUB domain
39	Ig C domain
39	prion repeat
37	ankyrin repeat
37	thioredoxin domain
33	GATA-type zinc finger
29	Zn-dependent phorbol ester/DAG binding
26	glucagon peptide
25	C2 domain
25	albumin domain
24	SH2 domain
24	HMG DNA-binding motif
23	vacuolar ATPase cytoplasmic domain
23	AAA-type ATP-binding
22	kringle domain
22	zein repeat
21	apoliprotein repeat
20	integrin b cysteine-rich repeat
20	phytochrome repeat
20	integrin a repeat
20	protein tyrosine-phosphatase domain
20	TFIID repeat
20	armadillo repeat
......	
2213 families total	

Prokaryotic Repeat Families

# Proteins	Description
45	ATP-binding cassette
29	glycine-rich calcium-binding motif
20	gyrase C-term DNA-binding domain
19	S-layer binding motif
18	bacterial 4Fe-4S cluster
18	DnaJ substrate binding motif
12	*S.pyogenes* M antigen coiled coil motif
12	cell wall binding domain
11	chemotaxis protein methylated domain
11	general secretion pathway protein domain
10	cellulase proline-threonine box
10	ribosomal S1 motif
10	drug efflux pump duplication
......	
861 families total	

Figure 4. Families of repeats. While the majority of repeats are relatively distinct sequences, many repeats can be clustered into homologous families. Shown here are the largest families ranked by prevalence for both eukaryotes and prokaryotes. Only families with 20 or more eukaryotic or ten or more prokaryotic sequence members are shown. Eukaryotic families related by homology to prokaryotic families are connected by lines; broken lines indicate that the related family has fewer than ten members.

greatest number of members (45 proteins) was the ATP-binding cassette family which is also found in eukaryotes. Some families included common repeating segments from unrelated proteins, such as glycine-rich calcium-binding motifs and [4Fe-4S] clusters. Other repeat segments were found only within clusters of related proteins, such as those from the C-terminal domain of topoisomerases.

In contrast to prokaryotic repeat segments, eukaryotic repeats cluster more extensively into families. Eukaryotic sequences were clustered into 86 large families holding 49 % of the sequences with repeats. The most populous eukaryotic repeat family is the zinc-finger motif (314 proteins), followed by calcium-binding motifs, collagen repeats, and WD repeats.

New classes of repeats

The clustering analysis has also led to the identification of several previously unobserved classes of repeats. Several of these are described in Table 1. The majority of the novel repeats are simple duplications, although several novel multiple tandem repeats are observed in both eukaryotes and prokaryotes. We anticipate that the novel families of tandem repeats will represent new protein folds, as did proteins containing armadillo repeats and leucine-rich repeats (Conti *et al.*, 1998; Kobe & Deisenhofer, 1993).

Mechanism of repeat formation

It has been suggested that the mechanisms underlying hypermutability of minisatellite loci (repeating units of more than ten nucleotides) are recombination events, while the evolution of the shorter microsatellites (repeating units of less than ten nucleotides) is caused by polymerase or strand slippage, possibly by formation of DNA hairpins (Kruglyak *et al.*, 1998; Buard & Vergnaud, 1994; Djian, 1998). Our census permits us to speculate about which of these mechanisms leads to the emergence of repeating sequences in proteins. In Figure 1(b) and (c), we examined the frequency of repeats as a function of protein length. To address the question of repeat-generating mechanisms we examined the frequency of repeats as a function of repeat length. The distribution on a linear-log plot (Figure 5(a)) reveals a linear trend, suggesting that the probability of generating repeats of a certain length decreases exponentially with length. One might expect the length-dependence to relate to the energy of forming the repeats. For example, if the mechanism of repeat formation involves DNA slippage or hybridization, one might expect the length dependence to relate to the energy per base-pair required to melt DNA. Assuming that the energy change associated with the repeat formation is proportional to the length of the fragment, we can adopt the simple model that the probability of forming a repeat is given by the Boltzman probability:

Table 1. Novel protein repeat families

Family description	Example[a]	Number of proteins	Size of repeats[b]
A. *Eukaryotic families*			
Chlorophyll-binding proteins	cb11_lyces	56	2×18-53 aa
Neurophysins[c]	neu1_mouse	16	2×12-20 aa
Cyclic GMP stimulated phosphodiesterases	cn2a_rat	11	2×20-75 aa
Sodium/calcium exchanger cytoplasmic domains	nac1_human	9	2×93-111 aa
Hypothetical chloroplast genes	ycf3_maize	8	2×16-30 aa
30 kDa moth hemolymph/vitellogenic proteins	lp1_bommo	7	2×23-52 aa
Major vault proteins	mvp_human	4	4.5×53 aa
Major pollen antigens	mp51_phaaq	4	2-3×40-127 aa
Urea transporters	ut1_human	4	2×68 aa
Pupal cuticle proteins	cug1_tenmo	2	12×10 aa
B. *Prokaryotic families*			
General protein secretion inner membrane proteins	hofc_ecoli	11	2×117-168 aa
Acriflavin drug efflux pumps	acrf_ecoli	10	2×51-149
ATP synthase CF(0) B/B′ chain	atpx_anasp	10	2-5×11-26 aa
Rickettsia surface antigens	17kd_ricty	7	3-5×8 aa
Adhesin proteins	aida_ecoli	4	14-38×18-19 aa
Periplasmic proteases	degq_ecoli	6	2×32-53 aa
Chlamydia envelope virulence factor	om6c_chltr	6	2×50-95 aa
General protein secretion outer membrane proteins	gspd_ecoli	5	2×30-83 aa
Integral membrane sensor proteins	evgs_ecoli	4	2×199-240 aa
Cell wall-associated proteases	p2p_acpa	4	2×41-43 aa
Intimins and invasins	eae2_ecoli	4	2×67-99 aa
Flagellar assembly proteins	flih_coli	2	8-9×4 aa

[a] SWISS-PROT entry name.

[b] Number of repeats observed and lengths of single repeating units (given as a range of numbers of amino acids (aa) found for single repeat units from different proteins).

[c] Repeating sequences of lowest acceptable significance.

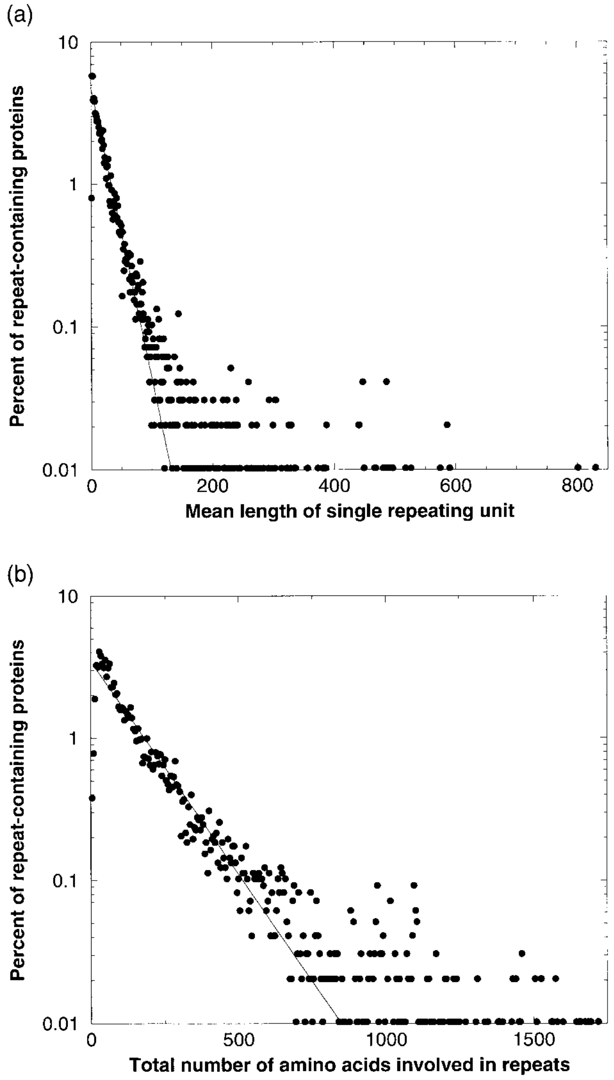


Figure 5. The size distribution of internal repeats in the proteins of SWISS-PROT 35.0. (a) The observed distribution of repeat lengths (filled circles) is not random. Instead, the length distribution is biphasic: repeats shorter than 150 amino acid residues in length fit an exponential (continuous line). (b) The observed distribution of the total number of amino acids participating in the repeating sequences is also biphasic: repetitive regions shorter than 800 amino acid residues fit an exponential curve (continuous line). Due to the logarithmic nature of the plots, data points in the lower right regions represent only a relatively small fraction of the proteins.

$$P_{\text{recomb}} \propto e^{-\frac{n\langle E\rangle}{kT}} \tag{1}$$

where $\langle E\rangle$ is the average energy of a nucleotide pair, and n is the fragment length. Fitting this model to the data in Figure 5(a) gives a value of $\langle E\rangle$ of 0.016 kT, on the order of 1/100 of the true melting energy per nucleotide pair (Fossella *et al.*, 1993). This argument suggests that the mechanism producing repeats (from five to ten to thousands of nucleotides) is far less sensitive to repeat length than would be expected if slippage and therefore duplex melting were the limiting factor. Instead, the result supports mechanisms such as recombination that show only weak length dependence.

Further information about the mechanism of repeat formation may be extracted by examining the distributions of the number of repeats. These distributions show a complex dependence on the repeat length (data not shown). The probability of repeat expansion is a function of both the repeat length and of the number of repeats. Shorter repeats are expanded at a higher rate, as are proteins that already contain repeats. However, the total number of amino acids participating in the repeats (that is, repeat length times number of repeats) shows a much simpler distribution, plotted in Figure 5(b). As with the repeat length distribution, the distribution of the total number of amino acid residues in repeats can be modeled by a simple Boltzman probability model, where the probability of repeat formation is a function of the total number of nucleotides involved. Fitting this model to the data in Figure 5(b) gives a value of $\langle E\rangle$ of 0.0023 kT, approximately one-tenth of the energy required to form the single repeating units of Figure 5(a), suggesting that the expansion of repeats is much easier than the initial repeat formation. As before, this distribution shows a weak length-dependence, consistent with a recombination mechanism for repeat formation. The distribution in Figure 5(b) diverges from the exponential fit at larger regions of repeats. This divergence can be interpreted physically as the tendency over the course of evolution for the repeat-producing machinery to generate additional repeating sequences with an efficiency depending upon the amount of preexisting repeating sequence. This tendency to duplicate larger regions of repeats preferentially mirrors the rapid pathological repeat expansion in diseases like Huntington's and fragile X syndrome, in which trinucleotide repeats expand with increasing probability as the number of repeats increases (Leeflang *et al.*, 1995; Reiss *et al.*, 1994).

Residues in repetitive sequences

Regardless of the underlying mechanism for repeat formation, the pressures of selection determine which repeating fragments are preserved in modern protein sequences. Before commenting on these effects we note that repeats can be naturally divided into two classes: "low-complexity" repeats that contain very non-uniform amino acid composition (e.g. runs of single amino acid or of similar amino acid residues) and "high-complexity" that are composed of longer repeat lengths with complex amino acid composition. The repeats we identified above were classified as either high or low-complexity based upon whether the sequences survive or are eliminated by a filter that identifies runs of single amino acids or closely related amino acid residues (Wootton & Federhen, 1993). Based on this classification, only 19 % of the repeating sequence fragments are of low-complexity.

In general, we find that repeats show significant deviations from the normal amino acid composition. In the high-complexity repeats, small and/or polar amino acids (proline, glycine, serine, glutamine, glutamate, and histidine) are over-represented and large and/or non-polar amino acid residues (leucine, lysine, cysteine, isoleucine, tryptophan, phenylalanine, alanine, tyrosine, and valine) are under-represented. The low-complexity repeats show a similar, but considerably more exaggerated trend, with elevated glycine, proline, serine, glutamine, and alanine and depressed occurrences of arginine, cysteine, methionine, tryptophan, phenylalanine, valine, isoleucine, and leucine. Each of the above lists is in descending order, ranked by change from amino acid composition of proteins in SWISS-PROT. This ranking of amino acids by deviation from SWISS-PROT composition is strongly correlated (probability of occurring by chance $= 7 \times 10^{-4}$) to the ranking by hydrophobicity divided by amino acid volume, suggesting that these physical properties are the basis of a selective pressure which produces the observed amino acid distribution in the repeats. The amino acids preferred in repeats are similar to the amino acids often found in loops and are more soluble and less bulky than amino acids in protein cores.

Discussion

Emergence of repetitive sequences

The sequence comparison of eukaryotic and prokaryotic repetitive sequences showed only a small (4 %) overlap between the eukaryotic and prokaryotic sequences. Most of the conserved fragments between the two superkingdoms are ATP-binding cassettes. Similarly, the primary conclusion from the clustering analysis is that prokaryotic and eukaryotic repeat families are, with few exceptions, not homologous to each other. We suggest therefore that the vast majority of repeating sequences emerged after the eukaryotic-prokaryotic divergence.

Evolution of repeats

We have shown that eukaryotes, far more than prokaryotes, have evolved repetitive proteins to perform functions specific to their unique physiological need. However, another intriguing possibility that might account for the abundance of repeat-containing proteins in eukaryotes has to do with evolutionary rates. It is well-established that the formation of non-protein-encoding repeating sequences is an error-prone process, with mutations in both genomic DNA micro- and minisatellite repeats occurring far more frequently than the background rate of point mutations (Kruglyak *et al.*, 1998; Buard & Vergnaud, 1994). This suggests to us that repetitive proteins may evolve more quickly than non-repetitive ones. Certain prokaryotes are believed to rely upon the variable nature of genomic repeats to generate novel surface antigens and thereby adapt to changing environments (Moxon *et al.*, 1994; Tomb *et al.*, 1997). Therefore, eukaryotic genomes, possibly compensating for longer generational times, may take advantage of this extra source of variability during evolution by coding for a greater number of repetitive proteins.

Some evidence to support this notion comes from the examination of repeat numbers within specific families of proteins, such as proteins with leucine-rich repeats. We find that the number of copies of leucine-rich repeats in a single protein varies from one to over 40 copies and has a nearly uniform distribution up to about 20 copies. This trend is quite common among tandem-repeat-containing protein families. In collagen proteins, the number of repeats varies from less than 100 copies to over 500 copies. This high level of variability indicates that the numbers of repeats are changing rapidly over the course of evolution.

Conclusions

We find repetitive sequences are more common in eukaryotic proteins than in prokaryotic proteins. These repetitive sequences apparently formed after the prokaryotic-eukaryotic divergence by a mechanism with weak length-dependence such as recombination. We suggest that repetitive proteins evolve quicker than non-repetitive proteins. The simplest assumption is that similar repeat-forming mechanisms are operating in the different superkingdoms to generate repeats, but eukaryotes possess a relatively sophisticated protein synthesis machinery that includes cytosolic and bound ribosomes, an endoplasmic reticulum and golgi, and glycosylation and cytosolic formation of disulfide bonds. This machinery is likely to provide eukaryotes an advantage over prokaryotes in handling the multi-domain, non-globular folds likely to be found among repeat-rich proteins. We suggest that eukaryotes use this advantage by incorporating more repeating sequences in their proteins, thereby gaining the benefits offered by repeats: modular construction of new proteins and introduction of rapidly evolving protein sequences which allow faster adaptation to new environments.

Materials and Methods

Each protein sequence was analyzed for the presence of repeats by aligning the amino acid sequence against itself and using a modified Smith-Waterman alignment algorithm (Smith & Waterman, 1981; Waterman & Eggert, 1987) to find the best sub-optimal (off-diagonal) alignment that does not intersect the diagonal. The method was repeated iteratively, disallowing previously calculated alignments by modifying the matrix of optimal partial scores to eliminate previously identified paths, to find all statistically significant sub-optimal alignments. This method is extremely rapid, scaling roughly as N^2, where N is the length of the sequence. Therefore, the algorithm is easily applied to thousands of sequences. Also, the method allows calculations of the statistical significance of the results using Poisson statistics to calculate the probability of obtaining a given sub-optimal alignment from the distribution of scores of random alignments using the method by Waterman & Vingron (1994). It makes no prior assumptions about the repeat lengths. Suboptimal alignments were considered statistically significant if they had a probability $p < 1 \times 10^{-3}$ of occurring by random chance, a threshhold that gives <10 % false positive rate for proteins in SWISS-PROT. To extract approximate repeat lengths and number of occurences, we analyzed the protein's suboptimal alignment path matrix, the $N \times N$ matrix where each entry is one if a statistically significant path passes that entry and zero otherwise. Projection of the path matrix onto one-dimension gives a step function whose steps correspond approximately to sequence repeats. Details of the suboptimal alignment method and the algorithm used to extract repeat lengths and number of occurrences from the sub-optimal alignments are described by Pellegrini *et al.* (1999). Protein sequences were selected from SWISS-PROT Release 35.0, from 15 published genomes available in 1998, and from the unpublished genome of *Pyrobaculum aerophilum* (Fitz-Gibbon, 1998). Proteins can be analyzed for repetitive sequences using these algorithms at http://www.doe-mbi.ucla.edu/people/matteo/repeats.html.

Analysis of internal amino acid repeats requires unusual handling of sequence homologs. For other sequence analyses, one might correct for multiple occurrences of similar sequences. However, we do not correct for multiple occurrences of homologs of a given sequence, because (1) we are analyzing internal amino acid sequences that often fall into otherwise unrelated protein sequences; and (2) the number of repetitive sequences often varies widely from homolog to homolog.

Amino acid sequences participating in repetitive sequences were identified from the highest-scoring suboptimal alignment of each protein in SWISS-PROT shown to have a statistically significant suboptimal alignment. The repetitive sequences were compared to one another using the Smith-Waterman alignment algorithm. The statistical significance of alignments were described by the probability P of obtaining a higher alignment score using shuffled sequences. Alignments were considered significant if they occurred with a probability P greater than a P value threshhold equal to $1/T$, where T is the total number of sequence comparisons performed. For example, when we compare n eukaryotic repetitive sequences to m prokaryotic sequences, we set $T = n \times m$. The sequences were clustered into families of homologous sequences by first performing all pairwise sequence alignments and then exhaustively clustering all sequences with statistically significant alignments. Any two sequences within a cluster are linked by virtue of direct homology or indirectly by homology to other sequences within the cluster. Sequences from different clusters have no direct or indirect homology.

For the purposes of automatic functional classification of repetitive sequences, we grouped proteins using the standardized keyword annotation of the SWISS-PROT database.

Sequences were analyzed for low or high-complexity regions using the method by Wootton & Federhen (1993). Sequences were classified as low-complexity if they were removed by the 'seg' filter implemented in the Wisconsin Package Version 9.1, Genetics Computer Group (GCG), Madison, Wisc, using a low-cutoff of 1.5 and high-cutoff of 2.2 (empirically derived settings that filter out prion poly(A) sequences but pass the tandem repeats.)

For the purpose of analyzing residues in repetitive sequences, values for amino acid hydrophobicities came from Radzicka & Wolfenden (1988), (units of kJ/mol for partitioning of side-chain analogs from water into cyclohexane). The value for proline was derived by interpolating between adjacent values for the partition coefficients of N-acetyl amino acid amides between water and octanol (Fauchere & Pliska, 1983). For volumes, Van der Waals volumes of amino acids were used. Ordered amino acid lists were compared by calculating an error value based on the differences in ranking between the lists:

$$S = \sum_{AA} |rank_{list1} - rank_{list2}| \qquad (2)$$

The probability of achieving a lower score by chance was calculated numerically from the distribution of scores for 100,000 randomized lists.

Acknowledgements

This work was supported by a Department of Energy (DOE)/Oak Ridge Institute for Science and Education Hollaender postdoctoral Fellowship (E.M.), a Sloan Foundation/DOE postdoctoral fellowship (M.P.), and grants from the DOE. We acknowledge many helpful discussions with Drs Robert Grothe and Michael Thompson.

References

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucl. Acids Res.* **27**, 260-262.

Brenner, S. E., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene duplications in *H. influenzae*. *Nature,* **378**, 140.

Buard, J. & Vergnaud, G. (1997). Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13**, 3203-3210.

Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell,* **94**, 193-204.

Djian, P. (1998). Evolution of simple repeats in DNA and their relation to human diseases. *Cell,* **94**, 155-160.

Fauchere, J. & Pliska, V. (1983). Hydrophobic parameters π of amino acid side-chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369-375.

Fitz-Gibbon, S. T. (1998). Whole genome sequence of the hyperthermophilic archaeon *Pyrobaculum aerophilum*, PhD dissertation, University of California, Los Angeles.

Fossella, J. A., Kim, Y. J., Shih, H., Richards, E. G. & Fresco, J. R. (1993). Relative specificities in binding of Watson-Crick base-pairs by third strand residues in a DNA pyrimidine triplex motif. *Nucl. Acids Res.* **21**, 4511-4515.

Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archael genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Heringa, J. (1998). Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.* **8**, 338-345.

Heringa, J. & Argos, P. (1993). A method to recognize distant repeats in protein sequences. *Proteins: Struct. Funct. Genet.* **17**, 391-411.

Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids Res.* **27**, 215-219.

Kachroo, P., Ahuja, M., Leong, S. A. & Chattoo, B. B. (1997). Organisation and molecular analysis of repeated DNA sequences in the rice blast fungus *Magnaporthe grisea*. *Curr. Genet.* **31**, 361-369.

Kobe, B. & Deisenhofer, J. (1993). Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature,* **366**, 751-756.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* **25**, 619-637.

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA*, **95**, 10774-10778.

Leeflang, E. P., Zhang, L., Tavare, S., Hubert, R., Srinidhi, J., MacDonald, M. E., Myers, R. H., de Young, M., Wexler, N. S., Arnheim, N. & Gusella, J. F. (1995). Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum. *Human. Mol. Genet.* **4**, 1519-1526.

McLachlan, A. D. (1983). Analysis of gene duplication repeats in the myosin rods. *J. Mol. Biol.* **169**, 15-30.

Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24-33.

Netzer, W. J. & Hartl, F. U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature,* **388**, 343-349.

Pellegrini, M., Marcotte, E. M. & Yeates, T. O. (1999). A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins: Struct. Funct. Genet.* **35**, 440-446.

Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry,* **27**, 1664-1670.

Reiss, A. L., Kazazian, H. H., Jr, Krebs, C. M., McAughan, A., Boehm, C. D., Abrams, M. T. & Nelson, D. L. (1994). Frequency and stability of the fragile X premutation. *Human Mol. Genet.* **3**, 393-398.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F. & Peterson, S., *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature,* **388**, 539-547.

Waterman, M. S. & Eggert, M. (1987). A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197**, 723-728.

Waterman, M. S. & Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Stat. Sci.* **9**, 367-381.

Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comp. Chem.* **17**, 149-163.

Edited by J. M. Thornton

(Received 29 June 1999; received in revised form 23 August 1999; accepted 24 August 1999)