Abundance and Distributions of Eukaryote Protein Simple Sequences*

Kim Lan Sim‡ and Trevor P. Creamer§

Protein simple sequences are a subclass of low complexity regions of sequence that are highly enriched in one or a few residue types. Such sequences are common in transcription regulatory proteins, in structural proteins, in proteins involved in nucleic acid interactions, and in mediating protein-protein interactions. Simple sequences of 10 or more residues, containing ≥50% of a single residue type are surveyed in this work. Both eukaryote and prokaryote proteomes are investigated with emphasis on the eukaryotes. Very large numbers of such sequences are found in all organisms surveyed. It is found that eukaryotes possess far more simple sequences per protein than do the prokaryotes. Prokaryotes display a linear relationship between number of proteins containing simple sequences and proteome size, whereas it is not clear that such a relationship holds for eukaryotes. Strikingly, it is found that each eukaryote possesses its own unique distribution of simple sequences. Within those distributions it is found that simple sequences enriched in certain residue types are clearly favored, whereas others are just as clearly discriminated against. The preferences observed are not correlated with residue occurrence. An analysis of classes of proteins of known function suggests that simple sequence occurrence and distribution may be related to protein function. Based upon this analysis, the large number of simple sequences found above that would be expected from a simple statistical model, plus the known functional importance of numerous such sequences, it is postulated that eukaryotes have evolved to not only tolerate large numbers of simple sequences but also to require them. ***Molecular & Cellular Proteomics 1: 983–995, 2002.***

Protein simple sequences are stretches of sequence highly enriched in one or a few residue types. These sequences form a major subclass of low complexity sequences (1). Such sequences are common in transcription regulatory proteins where they are often enriched in glutamine, proline, or charged residues and tend to be highly conserved (2–5). Glutamine-enriched sequences are thought to be the most common simple sequences (5, 6) and have been associated with a number of human neurological disorders such as Huntington's disease (7–10). Proline-rich sequences are known to have important roles as structural elements and in mediating protein-protein interactions (11, 12). Sequences enriched in charged residues have been associated with DNA and RNA processing, chromatin structure, ion binding, and protein-protein interactions (10, 13). Various simple sequences have been implicated as protein domain linkers (14) or as markers for disordered proteins (15, 16). Clearly there are numerous instances where such sequences play important functional roles. In addition, Kashi *et al.* (4) have noted that DNA simple sequences are a potential source of genetic variation. Some of these DNA sequences fall within coding regions, leading to variation at the protein level. The recent explosion in available genomic, and consequently proteomic data, has provided the opportunity to examine the occurrence and distribution of protein simple sequences at a level of detail not previously possible. Here we present a survey of the occurrence and distribution of protein simple sequences highly enriched in a single residue type in the proteomes of four eukaryotes whose genomes have been fully sequenced. The occurrence of eukaryote simple sequences is compared with the occurrence of such sequences in the proteomes of 26 prokaryotes.

Some previous studies of protein simple sequences have used somewhat limited protein databases and have not necessarily compared organisms (5, 6, 17, 18). Other surveys have considered whole proteomes but often remove sequences considered redundant (19, 20). There are a number of surveys where simple sequences enriched in a particular residue type or associated with a particular function have been examined (2, 3, 9, 14). Some recent studies have focused on comparisons between organisms (10, 21–23) but have mostly considered only homopolymeric sequences. Our current study differs from prior work in that we use only intact proteomes from fully sequenced genomes, including sequences annotated as hypothetical proteins. We focus solely on non-overlapping simple sequences, of 10 or more residues in length, highly enriched in a single residue type (≥50% composition). This approach provides a non-biased view of the distribution of this set of protein simple sequences as well as allowing for ready comparison of their occurrence in the organisms examined. The eukaryotes surveyed, namely a yeast, worm, fruit fly, and plant, comprise a diverse sample of members of the eukaryote kingdom. We have chosen not to include the human proteome given the current uncertain state of its completion. In addition, for comparison we have sur-

From the Center for Structural Biology, Department of Molecular and Cellular Biochemistry, University of Kentucky, Lexington, Kentucky 40536-0298

Received, May 24, 2002, and in revised form, October 16, 2002
Published, MCP Papers in Press, December 9, 2002, DOI 10.1074/mcp.M200032-MCP200

veyed 26 prokaryotes, including 12 Archaea, two cyanobacteria, and six Gram-negative and six Gram-positive bacteria.

We find that highly enriched simple sequences are remarkably common in all of the organisms examined. Eukaryotes are found to possess more simple sequences per protein than do the prokaryotes in keeping with the findings of other groups (19, 21, 23). The occurrence of prokaryote proteins containing simple sequences is linearly correlated with proteome size. Given the limited number of organisms examined, it is not clear that this is the case for the eukaryotes. Perhaps most notably, each organism examined possesses its own unique distribution of simple sequences. We find that simple sequences display surprising length dependences with some residues preferentially populating long simple sequences regions, while others clearly prefer short simple sequences. There is no discernible correlation with residue occurrence. For example, leucine-enriched sequences appear to be discriminated against despite leucine being the most common residue in most organisms. Some observed length dependences can be explained in structural and functional terms, although many remain enigmatic. We have also found that simple sequence distributions vary according to functional groupings. For example, leucine-rich regions, despite being discriminated against in the overall distributions, are among the most common simple sequences found in membrane-associated proteins. It is clear from the sheer number found that all organisms examined, particularly eukaryotes, tolerate, and perhaps even require, large numbers of protein simple sequences. The data presented here will provide the basis for future studies of these ubiquitous and potentially extremely important sequences.

EXPERIMENTAL PROCEDURES

Complete proteomes from the fully sequenced genomes of four eukaryotes and 26 prokaryotes were used in our studies (Table I). Sequences were obtained as FASTA format files from the European Bioinformatics Institute (www.ebi.ac.uk/genomes/). We use the entire proteome for each organism, including all proteins marked “hypothetical,” “putative,” or “probable” as well as all proteins that have no annotation. The one exception to this is the proteome of *Arabidopsis thaliana* (AT),[1] in which 782 of the protein sequences were found to be incomplete (3% of the proteome). We therefore used only the 26,496 complete sequences in the AT proteome.

We arbitrarily define simple sequences as stretches of sequence that 1) are at least 10 residues in length, 2) are composed of ≥50% of a single type of residue, 3) begin and end with the residue of interest, and 4) do not possess gaps (runs *without* residue of interest) of more than 5 residues in length.

We represent a protein sequence of length L as a string, $a_1a_2a_3a_4 \ldots a_L$, where a_i is the residue at position i. When searching for a simple sequence enriched in a certain residue type, the numerical positions in the protein string for that residue are first generated as a string of i values. Putative simple sequences are extracted based on the positions of the i values given that gaps of 6 or more residues in length are not allowed within a simple sequence. Putative simple sequences of many lengths are identified with all i values corresponding to the residue of interest being output. Since only the residue of interest is selected, the process automatically generates only sequences that begin and end with the residue of interest. Subsequent filtering removes sequences that are less than 10 residues long. Remaining sequences are tested to satisfy the ≥50% threshold for the residue of interest. Sequences that do not satisfy the criteria are further analyzed to determine whether shorter simple sequences satisfying our criteria are within them. The entire process results in the identification of all non-overlapping simple sequences within the proteomes that satisfy all four of the above criteria. The computer programs used to identify simple sequences were written in Python/C++ and executed on a Silicon Graphics work station.

We use the Poisson distribution (9, 24) to model the probability of random occurrence of simple sequences containing a given residue type in the eukaryote proteomes. This is given by

$$f(n) = \frac{e^{-m}m^{n}}{n!} \qquad \text{(Eq. 1)}$$

where $f(n)$ is the probability of an event happening n times. In our studies l is the length of the simple sequence, n is the threshold value, and m is derived from

$$m = \frac{l \times (\text{\% occurrence of residue})}{100} \qquad \text{(Eq. 2)}$$

The expected number of simple sequences of length l in a proteome is then

$$\mathrm{SS}_{\mathrm{expect}} = f(n) \times T_l \qquad \text{(Eq. 3)}$$

where T_l is the total of number of sequence windows of length l in the proteome.

The difference between the actual number of simple sequences, $\mathrm{SS}_{\mathrm{Tot}}$, of length l found and the number expected from the Poisson distribution is then

$$\Delta = \mathrm{SS}_{\mathrm{Tot}} - \mathrm{SS}_{\mathrm{expect}} \qquad \text{(Eq. 4)}$$

For simple sequences longer than about 25 residues, $\mathrm{SS}_{\mathrm{expect}}$ is essentially zero in which case Δ is equal to the number of simple sequences found. Finally, to compare the occurrence of simple sequences among organisms, we define Δ_R as follows:

$$\Delta_R = \frac{\Delta}{\text{Number of proteins in proteome}} \qquad \text{(Eq. 5)}$$

RESULTS AND DISCUSSION

Simple Sequence Definition

Our criteria for identifying protein simple sequences ensures that we find sequences that would satisfy any definition of simple sequences, such as the low complexity measures of Wootton and co-workers (1) or the definition used by Golding (20). We chose to use this definition since it is relatively straightforward to apply, and the sequences identified are unambiguous in nature. The allowable gap used (5 or fewer

[1] The abbreviations used are: AT, *A. thaliana*; AF, *A. fulgidus*; AgT, *A. tumefaciens* C58; AP, *A. pernix* K1; BH, *B. halodurans*; BM, *B. melitensis* 16M chr1; BS, *B. subtilis*; CA, *C. acetobutylicum* ATCC824; CE, *C. elegans*; DM, *D. melanogaster*; DR, *D. radiodurans* chr1; EC, *E. coli* K-12; HI, *H. influenzae*; HP, *H. pylori* 26695; HS, *Halobacterium* sp. NRC-1; MG, *M. genitalium*; MJ, *M. jannaschii*; MP, *M. pneumoniae*; MT, *M. thermoautotrophicum*; Nos, *Nostoc* sp. PCC7120; PA, *P. abyssi*; PAe, *P. aerophilum*; PH, *P. horikoshii*; SC, *S. cerevisiae*; SS, *Synechocytis* sp. PCC6803; SSol, *S. solfataricus*; ST, *S. tokodaii*; TA, *T. acidophilum*; TV, *T. volcanium*; VC, *V. cholerae* chr1.

residues) was chosen because this is the largest gap possible in a 10-residue sequence, the shortest considered, while still satisfying our ≥50% threshold requirement. The ≥50% threshold ensures that even the shorter sequences identified are relatively unlikely to have occurred as a result of randomness in protein sequences. As will be demonstrated below, for many residues the Δ values obtained tend to be large and positive, indicating that we did indeed identify many more sequences than would be expected were sequences random in nature. If the threshold is decreased to ≥30%, we find significantly more simple sequences at all lengths; however, many of these, particularly short sequences, are accounted for by the number expected using the Poisson distribution model (data not shown). If the threshold is increased to ≥70% we find relatively few sequences (data not shown).

Inclusion of Potentially Incorrect Protein Sequences

We have chosen to include all complete protein sequences in the proteomes that we have examined. This includes those marked hypothetical, putative, or probable and those proteins that have not as yet been annotated. Redundant sequences have also been included. This choice was made so as to be able to perform a more complete analysis of the proteomes, leading to an "unbiased" view. It is possible that some of the simple sequences found come from sequences that are not expressed as proteins. Bork and Copley (25) have pointed out that the identification of genes in sequenced genomes is difficult. It is particularly difficult for eukaryote genes where the identification of exons is error-prone. Ideally the analyses presented below should be repeated leaving out those proteins marked hypothetical or not annotated. This is, however, extremely difficult due to the wide variety of annotations used to denote such putative protein sequences. We have thus chosen to present the analyses of the complete proteomes with the caveat that some of the results may be slightly skewed by the presence of incorrect protein sequences.

Abundance of Protein Simple Sequences

All of the organisms surveyed possess a remarkable number of simple sequences in their proteomes (Table I). The number found ranges from 251 in the small proteome of MG (480 proteins) up to 27,542 in the proteome of AT (26,496 protein sequences surveyed). Furthermore, a remarkable fraction of proteins in each proteome possess at least one simple sequence. Fig. 1*a* is a plot of the number of proteins possessing one or more simple sequences, Prot_{SS}, against the number of proteins in each proteome. At first glance one might deduce that there is a linear relationship between the number of simple sequence-containing proteins and the total number of proteins. The line of best fit drawn in Fig. 1*a* has a correlation coefficient of 0.99. However, the eukaryotes possess significantly larger proteomes than do the prokaryotes and consequently far more simple sequences. In effect, the fit to the data is reduced to a fit to five points, the four eukaryotes plus the prokaryotes essentially as a single point.

If one considers just the four eukaryotes surveyed, a line of best fit through the data in Fig. 1*a* would yield a correlation coefficient of 0.99. Note, however, this is just a four-point fit and that it may well be that there is not a linear relationship between eukaryote proteome size and Prot_{SS}. Clearly the complete proteomes of more eukaryotes need to be examined, once they become available, to gain a better understanding of this relationship. What can be concluded from this figure, and the data in Table I, is that a remarkable number of the proteins in the eukaryote proteomes surveyed possess at least one simple sequence as defined in this work. The individual amounts are 53% of the proteins in SC, 51% in CE, 59% in DM, and 55% in AT. Why DM would possess a significantly higher fraction of proteins with at least one simple sequence is not clear. Karlin *et al.* (10), in a recent survey of homopolymeric runs in proteins ≥200 residues in size, found that DM possessed far more than other eukaryotes. They also found that human proteins possessed more of these runs than proteins from CE despite there being more CE proteins surveyed. Such data suggest that the human proteome may also possess a larger fraction of proteins containing simple sequences than the average observed in this work.

Fig. 1*b* is a plot of Prot_{SS} against the number of proteins in each proteome for the 26 prokaryotes surveyed. There is a clear linear correlation with the line of best fit having a correlation coefficient of 0.92. Two prokaryotes, the Archaea HS and the bacteria DR, appear to be outliers. Excluding these from the fit results in a correlation coefficient of 0.96. The strong linear correlation observed for the prokaryotes might suggest that these simple sequences have arisen via random events, leading to random distributions that depend only upon the number of proteins in each proteome. As will be demonstrated below, however, our data suggest the opposite, that the occurrence and distributions of simple sequences is not random in nature and that many of these sequences may possess biological significance.

Fig. 1*c*, a bar plot of the ratio of number of simple sequences found, SS_{Tot}, to Prot_{SS} for each organism surveyed illustrates the difference in occurrence of protein simple sequences in prokaryotes and eukaryotes. Prokaryotes have far fewer simple sequences per protein than do the eukaryotes. In all cases, the prokaryotes have fewer simple sequences than the total number of proteins in their proteomes, whereas the eukaryotes possess more (Table I). The prokaryotes average 1.40 simple sequences per protein possessing at least one simple sequence (the *dashed line* on Fig. 1*c*). Once again, HS and DR are clear outliers among the prokaryotes, possessing $\text{SS}_{\text{Tot}}/\text{Prot}_{\text{SS}}$ ratios of 1.68 and 1.73, respectively, both values greater than 2 standard deviations from the mean for prokaryotes. The eukaryotes have ratios that range from 1.88 in AT through 2.09 in CE and 2.18 in SC up to 3.09 simple sequences per protein possessing at least one simple se-

TABLE I

Organisms surveyed for protein simple sequences, the number of proteins in each proteome, total number of simple sequences found (SS_{Tot}), and the number of proteins containing at least one simple sequence ($Prot_{SS}$)

Organism	Two-letter code	Type	Number of proteins in proteome	SS_{Tot}	$Prot_{SS}$	$SS_{Tot}/Prot_{SS}$
Saccharomyces cerevisiae	SC	Eukaryote	6,203	7,177	3,293	2.18
Caenorhabditis elegans	CE		21,962	23,295	11,125	2.09
Drosophila melanogaster	DM		13,608	24,725	7,989	3.09
A. thaliana[a]	AT		26,496	27,542	14,637	1.88
Synechocytis sp. PCC6803	SS	Cyanobacteria	3,169	1,493	1,034	1.44
Nostoc sp. PCC7120	Nos		5,368	2,497	1,762	1.42
Escherichia coli K-12	EC	Gram-negative bacteria	4,289	2,064	1,636	1.26
Haemophilus influenzae	HI		1,709	614	476	1.29
Vibrio cholerae chr1	VC		2,736	1,219	881	1.38
Helicobacter pylori 26695	HP		1,566	699	500	1.40
Brucella melitensis 16M chr1	BM		2,059	1,067	756	1.41
Agrobacterium tumefaciens C58	AgT		2,722	1,610	1,078	1.49
Bacillus subtilis	BS	Gram-positive bacteria	4,367	1,723	1,270	1.36
Bacillus halodurans	BH		4,066	1,597	1,182	1.35
Mycoplasma pneumoniae	MP		688	360	242	1.49
Mycoplasma genitalium	MG		480	251	170	1.48
Deinococcus radiodurans chr1	DR		2,579	2,274	1,311	1.73
Clostridium acetobutylicum ATCC824	CA		3,672	1,550	1,132	1.37
Archaeoglobus fulgidus	AF	Archaea	2,421	990	747	1.32
Aeropyrum pernix K1	AP		2,694	1,827	1,176	1.55
Methanobacterium thermoautotrophicum	MT		1,869	691	531	1.30
Methanococcus jannaschii	MJ		1,715	875	641	1.36
Pyrococcus abyssi	PA		1,765	847	613	1.38
Pyrococcus horikoshii	PH		2,064	973	730	1.33
Halobacterium sp. NRC-1	HS		2,058	1,785	1,060	1.68
Thermoplasma acidophilum	TA		1,478	511	395	1.29
Thermoplasma volcanium	TV		1,526	452	376	1.20
Pyrobaculum aerophilum	PAe		2,605	1,220	866	1.41
Sulfolobus tokodaii	ST		2,826	1,203	866	1.39
Sulfolobus solfataricus	SSol		2,994	1,335	962	1.39

[a] Some AT protein sequences were incomplete and were not included in the analysis. The number of proteins listed for AT corresponds to the number used.

quence in DM. Eukaryotes clearly not only tolerate a significantly higher occurrence of these sequences than do the prokaryotes, they are also more likely to possess multiple simple sequences in each protein.

The ratio $SS_{Tot}/Prot_{SS}$ is of course dependent upon our definition of protein simple sequences. One can imagine that increasing the size of the allowable gap (currently set at 5 or fewer residues) will result in some of the simple sequences merging, resulting in fewer overall but an increase in the number of longer sequences. The result will be lower values of $SS_{Tot}/Prot_{SS}$ for each proteome.

A number of groups have examined the occurrence of homopolymeric runs of sequence and noted that eukaryotes possess more per protein than do prokaryotes (19, 21, 23). Nishizawa *et al.* (23) note that "modern" tissue-specific proteins have a higher tendency to possess homopolymeric stretches of up to 20 residues in length as compared with ancient proteins. They go on to postulate that this repetitiveness enhances the chance for intermolecular interactions. This hypothesis is supported by observations that simple sequences enriched in glutamine, proline, or charged residues are often found in protein interaction domains of transcription regulatory proteins (2–5) and that proline-rich sequences are common protein-protein interaction domains (11, 12). It seems likely then that eukaryotes, in particular the multicellular organisms, have evolved to require numerous protein simple sequences for functional purposes.

It is not clear why HS and DR would be outliers among the prokaryotes in Fig. 1. HS is an extreme halophile (26), the only one in the set of organisms surveyed. It is tempting to postulate that HS might possess a higher proportion of simple sequences as a result of evolving to survive in such an unusual environment. Ng *et al.* (26) pointed out that 36% of the putative proteins in the HS proteome were unrelated to any previously reported at that time and that these proteins may well provide the mechanisms by which HS can survive extreme salt concentrations. However, the HS proteome has not been analyzed in sufficient detail to know whether those proteins are particularly enriched in simple sequences, so we cannot draw any conclusions at this point.

DR has been nicknamed "Conan the bacterium" for its amazing ability to resist very high doses of ionizing radiation and UV irradiation (27) and is the only organism surveyed to possess these remarkable traits. It has been speculated that the radiation resistance of DR is due to its unique polypoid nature and the abundant DNA repeat elements in its genome. These DNA repeats may function to regulate DNA degradation after damage to this organism. The high number of protein simple sequences identified in this species may be attributed to such repeats although not to the polypoid nature of DR. This organism possesses more simple sequences *per protein* than do other prokaryotes (Table I). Simply possessing multiple copies of each gene would not raise the number of simple sequences per protein. The protein simple sequences may have arisen over time as a result of errors made by the DNA repair apparatus of DR while "rebuilding" its genome from multiple gene copies after exposure to extreme conditions such as radiation. On the other hand, some of these simple sequences may play an active role in the survival mechanisms developed by DR. Further functional analysis of the DR proteome is required to better understand why this organism possesses so many protein simple sequences.

For reasons of clarity and focus, the remainder of this article will focus on the occurrence and distributions of protein simple sequences in eukaryotes.

Overall Length Distributions

Fig. 2, a log-log plot of the number of simple sequences found against simple sequence length, is a clear illustration of the remarkable simple sequence length distributions observed in the four eukaryotes examined. Prokaryotes display similar length distributions, although generally the longest prokaryotic simple sequences are shorter than the longest eukaryotic sequences (data not shown). At the shorter simple sequence lengths a periodicity in the data can be seen with there being fewer occurrences where the length is an odd number as compared with adjacent even-numbered lengths. This is a consequence of the algorithm used to identify the simple sequences. As an example, given the threshold of ≥50%, a simple sequence 11 residues long must possess at least 6 residues of a given type. This amounts to a minimum of 55% enrichment, whereas a 12-residue simple sequence can also possess 6 residues, leading to a minimum of 50% enrichment. This periodicity tends to be damped out at long simple sequence lengths.

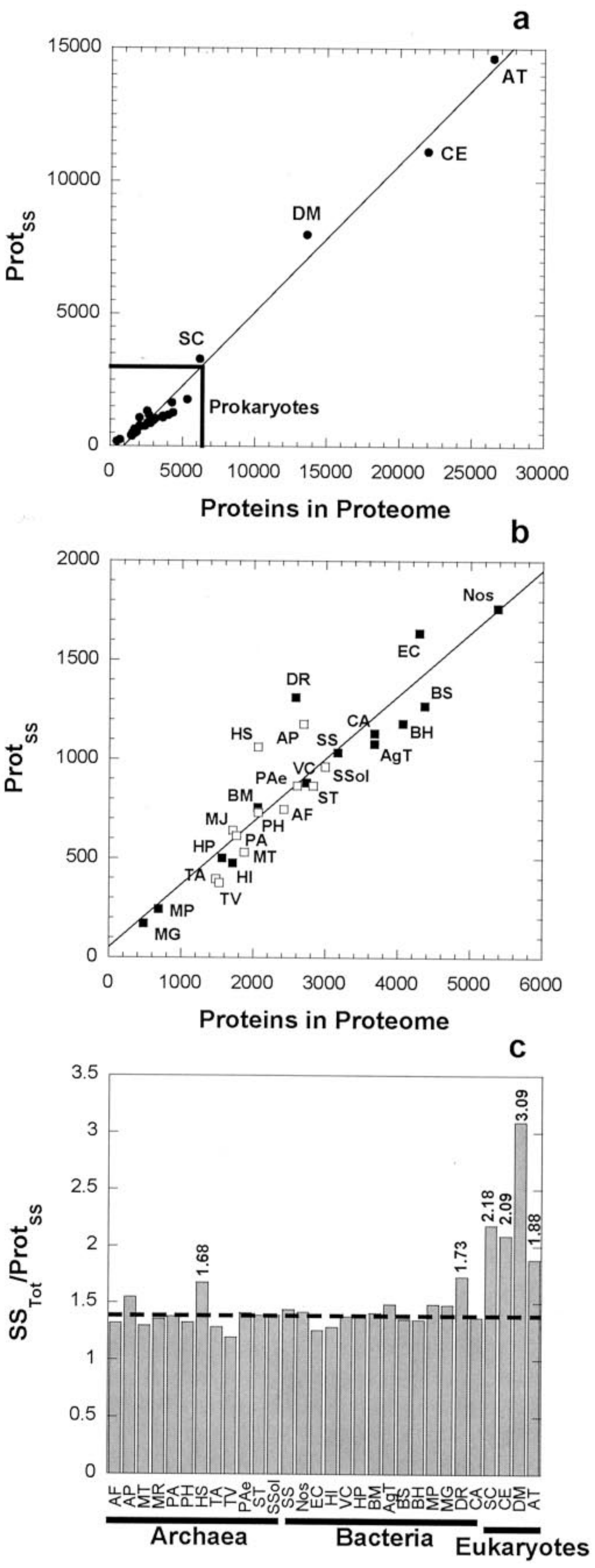

FIG. 1. *a* shows the number of proteins possessing at least one simple sequence plotted against the total number of proteins in the proteome for all organisms. *b* shows the same data for just the prokaryotes. Line of best fit was calculated excluding HS and DR. *c* is a bar plot of the ratio of simple sequences to the number of proteins possessing simple sequences for each organism. The *dashed line* denotes the average value for prokaryotes (1.40). The ratios for the eukaryotes and the two outlying prokaryotes (HS and DR) are provided.

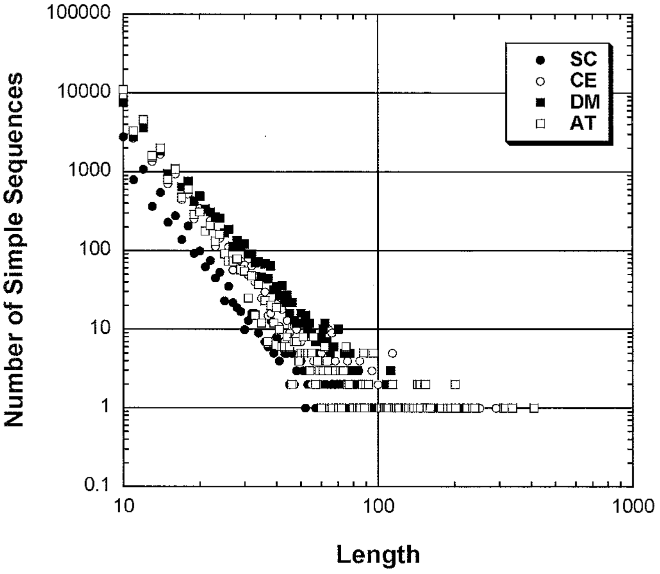

FIG. 2. **A log-log plot of the total number of simple sequences in each of the eukaryotes plotted against simple sequence length.**

Not surprisingly there is a sharp decrease in the number of simple sequences found with increasing length. The shorter simple sequences are extremely common. Of course such observations are in part a result of our definition of a protein simple sequence. Lowering or raising the threshold of ≥50% enrichment will change these numbers as will changing the gap between putative simple sequences. Nonetheless, protein simple sequences highly enriched in a single residue type are remarkably common.

The longest simple sequence was found in AT, is 410 residues long, and is enriched in glycine. AT is not alone in possessing remarkably long simple sequences. The longest in SC is 246 residues long and is enriched in serine. The longest in CE is threonine-rich and is 291 residues long, while DM possesses a 322-residue-long glycine-rich sequence. Notably, all four of these simple sequences occur in proteins that have been annotated as being hypothetical. The majority of the simple sequences found are of course much shorter than these, the vast majority being 60 or fewer residues in length (~99.5%; Fig. 2).

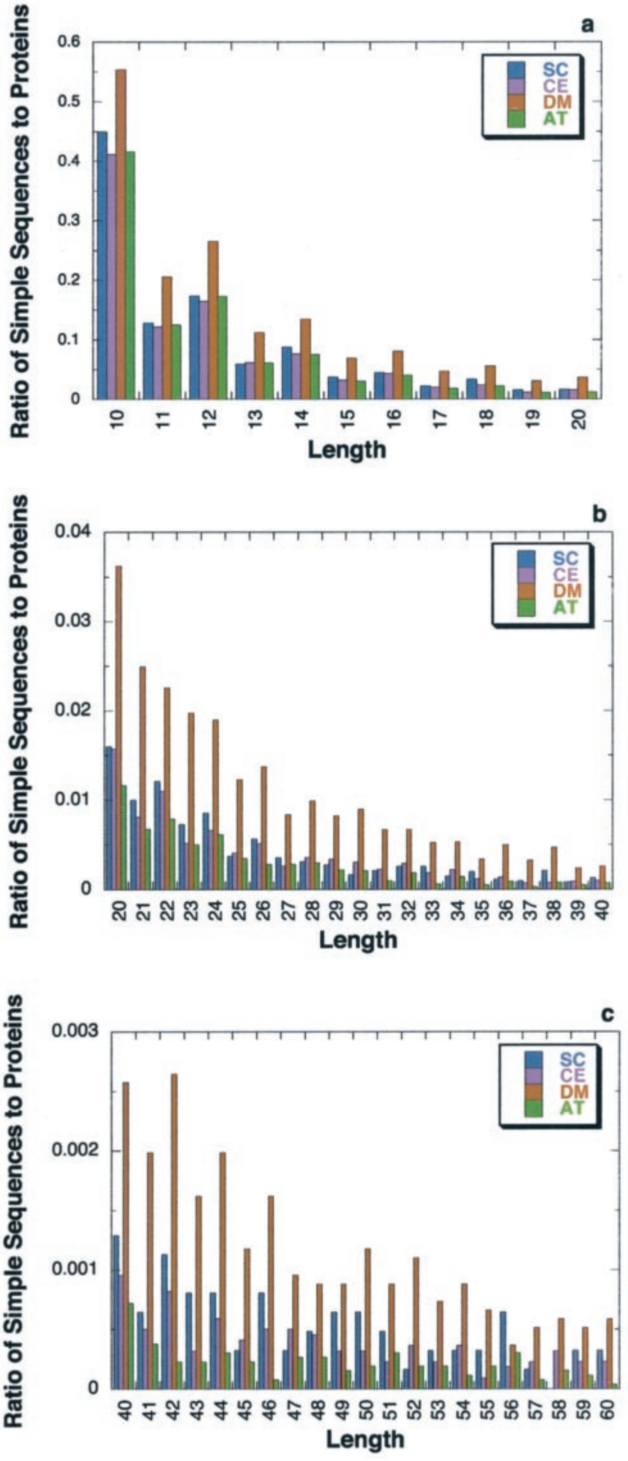

FIG. 3. **Ratio of total number of simple sequences to the number of proteins in each proteome for various simple sequence lengths.** *a* is data for simple sequences of 10–20 residues in length, *b* is for 20–40 residues, and *c* is for 40–60 residues.

Fig. 3 is a bar plot of the ratio of the number of simple sequences found to the number of proteins in the proteome as a function of length for the four eukaryotes. Division by the proteome size allows for direct comparison of the organisms. The data are split into three length scales; 10–20 residues (Fig. 3*a*), 20–40 (Fig. 3*b*), and 40–60 (Fig. 3*c*). The periodicity observed in Fig. 2 is obvious in Fig. 3*a* and can be seen to have subsided in Fig. 3*b*. It is clear from Fig. 3 that DM averages more simple sequences per protein at all lengths than do the other organisms despite AT possessing more in total and CE possessing a similar number (Table I). In fact, DM possesses more than twice as many simple sequences of lengths ≥20 residues per protein than do any of the other three eukaryotes examined. Clearly DM has evolved to tolerate large numbers of simple sequences. What is not entirely clear is whether this observation is linked to the functional requirements of DM. Nishizawa *et al.* (23) have pointed out that neural and immune system-specific proteins have a higher propensity to possess short runs of sequence consisting entirely of one residue type. One could reasonably expect that this would extend to the highly enriched simple sequences found in this survey. If so, it may not in fact be surprising that DM possesses such an abundance of these simple sequences as compared with the other eukaryotes examined. An analysis of the functions of the proteins in DM possessing simple sequences will shed light on this as will surveys of the proteomes of other eukaryotes as they become available.

What is perhaps most surprising in Fig. 3 is that SC possesses the second largest fraction of proteins in its proteome averaging a single simple sequence at almost all lengths up to 60 residues. AT generally possesses the fewest. Huntley and Golding (19) had previously noted that SC possesses a high proportion of protein simple sequences, although they could not explain why. The common theme from Fig. 3 is that many of the proteins in the proteomes of the four eukaryotes examined possess simple sequence regions. In fact, it has previously been observed that protein simple sequences are the most commonly shared sequence pattern among the eukaryotes (19). Huntley and Golding (19) have suggested that protein simple sequences are the equivalent of "junk DNA," serving little purpose. However, given the number of simple sequences found coupled with the known functions of some of them (2, 3, 10–12), it is tempting to postulate that eukaryotes tolerate and even require large numbers of simple sequences for functional reasons.

Residue Length Dependences

From Figs. 2 and 3 it would appear that the four eukaryotes examined have similar protein simple sequence distributions albeit with differences in relative abundance. Striking differences between the organisms are revealed when simple sequence distributions are considered at the level of individual residue types. Fig. 4 shows the ratio of the number of simple sequences found above that expected from the Poisson distribution to the number of proteins in each organisms proteome, Δ_R, plotted against simple sequence length for each residue. The sequence lengths are binned into ranges: 10–20 (Fig. 4*a*), 21–40 (Fig. 4*b*), and 41 and more (Fig. 4*c*) residues. Data for cysteine, methionine, and tryptophan are omitted since we found very few simple sequences containing these rare residues. The ratio Δ_R is a measure of how common simple sequences are, above the Poisson distribution predictions, per protein in each eukaryote proteome. This ratio allows for easy comparison of the organisms. A higher Δ_R indicates that simple sequences of a given length in an organism are more common in comparison to the other organisms even though the actual number found might be the same or even lower. A negative value of Δ_R indicates that those simple sequences are found less often than predicted from the Poisson distribution. Such sequences are presumably discriminated against for various reasons.

Features Common to All Eukaryotes Examined—Before considering differences between the distributions of simple sequences for each organism, there are some features common to all four eukaryotes worth looking at in Fig. 4. Perhaps the most obvious common features are the negative values of Δ_R at short lengths (10–20 residues) observed for the small apolar residues isoleucine, leucine, and valine (Fig. 4*a*). These negative values indicate that there are fewer such simple sequences than might be expected from the Poisson distribution. The most striking observation is that for leucine, the most common residue. We find hundreds fewer leucine-rich sequences at short lengths than expected. This is particularly apparent for SC and AT, which have Δ values of -585 and -1523, respectively. CE and DM also have large negative Δ values (-497 and -397, respectively). For simple sequences of 21–40 residues (Fig. 4*b*), the Δ_R values for leucine, isoleucine, and valine become positive but are small. For even longer lengths the Δ_R values are zero or very small. We appear to be observing a discrimination against simple sequences highly enriched in these small apolar residues. This has previously been observed by Green and Wang (6), Katti *et al.* (5), and Karlin *et al.* (10), who all found very few occurrences of runs of these residues longer than 10 residues in length. In these studies a run of residues was defined as consisting solely of a single type of residue except for in the study of Katti *et al.* (5) where a 10% mismatch was allowed for sequence runs greater than 20 residues in length. We may be observing a biophysical effect here. Sequences of 10–20 residues in length with $\geq$50% leucine, isoleucine, or valine will be highly hydrophobic and may pose an aggregation risk for proteins that contain them. Hence they are evolutionarily discriminated against. Moderate lengths, 21 to $\sim$30 residues become more likely since such sequences could act as membrane-spanning regions as suggested by Schwartz *et al.* (28).

We should note that the actual number of leucine-, isoleucine-, and valine-enriched simple sequences found can be quite large. For example, in DM we find 1841, 96, and 223 simple sequences of 10 residues in length enriched in each of these residues, respectively. Due to the relative abundance of these residues, however, the Poisson distribution predictions are also large (1445, 95, and 218, respectively), leading to small or negative values of Δ and Δ_R.

It is notable that we find positive values of Δ_R for phenylalanine and tyrosine at short, moderate, and even long lengths (Fig. 4). The tyrosine-rich sequences are particularly surprising given that this is one of the rarer residues. One might expect that sequences enriched in such large hydrophobic residues might be disfavored, and yet this does not appear to be the case. It is not clear why such sequences would be tolerated.

Careful inspection of Fig. 4 reveals that sequences highly enriched in serine, glutamate, lysine, and alanine appear to be favored by all four of the eukaryotes examined at short lengths

(Fig. 4*a*). At moderate lengths alanine-rich sequences become less common (Fig. 4*b*), while at long lengths glycine-rich sequences seem to be favored. Similar distributions of runs of sequence containing these residues were observed by Green and Wang (6) and Katti *et al.* (5), although these authors did not normalize their data for residue occurrence nor for what might be expected were sequences random in nature. It is not entirely apparent why sequences highly enriched in serine are tolerated, or even required, by eukaryotes, although there are certainly examples of important protein domains enriched in this residue. One such example is the C-terminal domain of RNA polymerase II, which is functionally essential and consists of the heptad YSPTSPS repeated between 26 and 52 times in various organisms (29). Interestingly, this serine-enriched region (~43% serine) is known to interact with proline-rich regions (12) as well as a family of serine/arginine-rich proteins (30). Wootton and Drummond (14) have suggested that sequences enriched in serine may act as flexible linkers between protein domains in much the same way as postulated for glycine-rich sequences.

Sequences enriched in charged residues, such as the lysine- and glutamate-rich sequences seen to be favored by the eukaryotes (Fig. 4), have been associated with DNA and RNA processing, chromatin structure, ion binding, and protein-protein interactions (13). The involvement of such simple sequences in a wide variety of functional roles might therefore explain their relative abundance. Alanine is known to be the most energetically favorable residue in α-helices (31, 32). One might therefore expect that sequences that are 50% or greater alanine in composition will have a tendency to be α-helical, although this will of course be modulated by the nature of the other residues in the sequence as well as by the tertiary structure of the proteins they are part of. The preference for short alanine-rich sequences that we have observed might then be related to secondary structure requirements. The long glycine-rich sequences found are probably tolerated for the opposite reason; that is, these most likely represent flexible linkers between protein domains.

One of the more surprising observations from Fig. 4 is that of simple sequences highly enriched in histidine. Although there are not many of these at all lengths, the number we find above that expected is significant. Some of these are quite long. For example, the four longest histidine-rich sequences in DM are 46, 51, 54, and 56 residues long. In CE the four longest are 50, 51, 84, and 251 residues long, although the longest of these is in a protein annotated as being hypothetical and could in fact be an indication that this is not an expressed protein. Histidine is one of the most rare residues, comprising just 2.2–2.7% of all residues in the four pro-

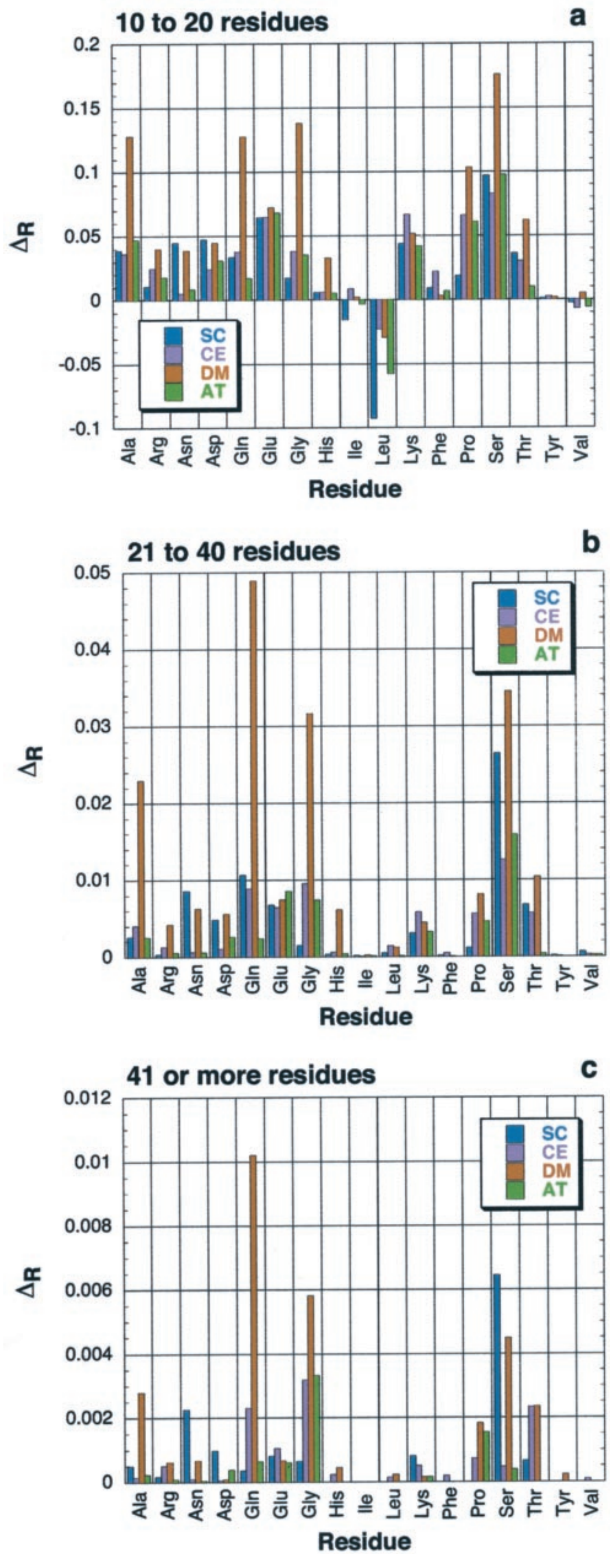

FIG. 4. **Ratio of the number of simple sequences found above that expected from the Poisson distribution model, Δ_R, to the number of proteins in each eukaryote proteome plotted for each residue type.** Data for cysteine, methionine, and tryptophan are excluded due to the low numbers of simple sequences found for these residues. *a* shows Δ_R for simple sequences 10–20 residues in length, *b* shows 21–40 residues, and *c* shows 41 or more residues.

teomes. By comparison, methionine has a similar level of occurrence, and yet we find almost no simple sequences enriched in this residue above the Poisson distribution predictions. Similarly, we find very few tryptophan- and cysteine-enriched sequences. One might postulate that histidine-rich sequences have some kind of ion binding function, although this has not been demonstrated.

Distribution Differences among the Eukaryotes Examined—It is immediately apparent from Fig. 4 that DM has a markedly different distribution of simple sequences when compared with the other three eukaryotes. The data for DM demonstrate a preference for simple sequences of all lengths enriched in alanine, glutamine, glycine, and serine. At short to moderate lengths, 10–40 residues (Fig. 4, *a* and *b*), DM also shows some preference for asparagine-, proline-, threonine-, and perhaps most surprisingly histidine-enriched sequences. To a lesser extent, there may also be a preference for aspartate- and arginine-rich sequences. These observed preferences are in large part responsible for the "unusually" high $SS_{Tot}/Prot_{SS}$ ratio observed for DM (Fig. 1*c*). The large numbers of glutamine- and to some extent asparagine-rich sequences in DM were also observed by Michelitsch and Weissman (9), who suggested that many of these may act as protein-protein interaction domains. It is not clear why DM would tolerate, and perhaps even require, large numbers of alanine-, glycine-, and serine-rich sequences.

Although DM is clearly different than the other three eukaryotes, it would be a mistake to assume that there are no significant differences between the distributions observed for the other organisms. SC has preferences for asparagine- and aspartate-enriched sequences at all lengths along with a striking preference for moderate length to long serine-rich sequences. Furthermore, SC disfavors leucine- and isoleucine-rich sequences more than do the other eukaryotes and is somewhat less tolerant of arginine-, glycine-, and proline-rich sequences. The reasons behind each of these preferences are not always clear. For example, the reasons for the large preference for moderate length to long serine-rich sequences are unknown. Wootton and Drummond (14) have suggested that sequences rich in serine form flexible linkers between protein domains. If this is true, then the preference for serine-rich sequences in SC may be linked to the observed lower tolerance for glycine-rich sequences (Fig. 4). SC may have evolved to use serine-rich sequences as linkers instead of the glycine-rich sequences that the other eukaryotes seem to prefer. Another potential role for serine-rich regions is discussed below. Michelitsch and Weissman (9) have previously observed large numbers of asparagine-rich sequences in SC as well as in other eukaryotes. These authors postulate that such regions act as modulators of protein-protein interactions. Why SC would require a larger fraction of asparagine-rich sequences for such interactions as compared with the eukaryotes is unclear. The lower tolerance for proline-rich sequences is probably due to the unicellular nature of SC. It has no need for the proline-rich extracellular structural proteins that the multicellular eukaryotes require. The reasons for the discrimination against leucine- and isoleucine-enriched sequences and the lower tolerance for arginine-rich sequences remain enigmatic.

The worm CE also possesses its own unique distribution of protein simple sequences. From Fig. 4 it can be seen that CE has some preference for short phenylalanine-rich sequences and for long glutamine- and serine-enriched sequences. CE also appears to be less tolerant of asparagine-rich sequences than are SC, DM, and perhaps AT and is less tolerant than are DM and AT of long proline-rich sequences. AT has little tolerance for threonine-rich sequences and a lowered tolerance for glutamine-rich sequences (Fig. 4). AT does not appear to have a heightened preference for any particular simple sequences at any length scale compared with the other eukaryotes.

It is clear that each of the four eukaryotes examined possesses its own unique distribution of simple sequences (Fig. 4). Based upon the analysis of homopolymeric runs performed by Karlin *et al.* (10) and an analysis by Kreil and Kreil (33) of asparagine-rich sequences, it seems clear that the human proteome will also display a unique simple sequence distribution. Some of the differences observed for the four eukaryotes examined arise for understandable reasons. For example, SC would not be expected to possess as many proline-rich sequences as would the other eukaryotes examined since SC does not have the same requirements for proline-rich structural proteins. However, as noted repeatedly above, the reasons for many of the various simple sequence preferences observed are not known. Some differences might well arise as a result of an organism using particular residues for the same purposes as other organisms use a different set of residues. For example, as suggested, SC might utilize serine-rich regions as flexible linkers where CE, DM, and AT use glycine-rich sequences. A detailed analysis of the conservation of simple sequence regions will aid in resolving such issues. Huntley and Golding (19) have noted that simple sequences are the most commonly shared feature between proteins but that the identity of the residues within the sequences can vary between organisms.

Functional Analysis of Protein Simple Sequence Occurrence—Our survey of the eukaryote (and prokaryote) proteomes has resulted in the identification of an enormous number of protein simple sequences, far more than would be expected were sequences random in nature. We have postulated that many of these sequences play some kind of functional role. This postulate is supported by a limited amount of experimental and bioinformatic evidence (3, 5, 9–12, 29, 30, 34). To further examine this issue we have examined the distribution of simple sequences in proteins of known function. Specifically, we have collected the sequences of all proteins from each of the four eukaryotes that are annotated in the SWISS-PROT database (35, 36) as being involved in a

TABLE II
Simple sequence distribution among proteins grouped according to class or process

Keyword[a]	SC		CE		DM		AT	
	Proteins	SS_{found}	Proteins	SS_{found}	Proteins	SS_{found}	Proteins	SS_{found}
Cell cycle	102	177	11	9	17	37	11	10
Metabolism	75	59	21	11	16	5	27	16
Signal	229	455	189	202	251	399	201	176
Transcription	274	677	87	101	177	838	105	130
Transport	440	449	148	85	102	133	160	118
Membrane	1,004	1,192	399	388	426	642	311	298

	Most common simple sequence type (number found)			
Cell cycle	Ser (64)	Ala (3)	Ser (11)	Ser (3)
Metabolism	Ala (13)	Gly (4), Leu (4)	Val (2)	Ser (6)
Signal	Ser (173), Thr (133)	Pro (34), Ser (24), Leu (24), Glu (24)	Leu (63), Ser (55), Gln (48), Ala (40)	Pro (61), Leu (42)
Transcription	Ser (138), Asn (114)	Ser (42)	Ser (194), Gln (161), Ala (136), Gly (99)	Ser (35), Ala (21)
Transport	Ser (88), Leu (75)	Ala (16), Leu (16)	Leu (22), Ala (21)	Ala (32), Ser (20)
Membrane	Ser (283), Leu (235), Thr (111)	Leu (77), Ser (47)	Leu (178), Ser (77), Gly (71), Ala (63)	Leu (85), Ala (45)

[a] Keyword used to search SWISS-PROT database for related proteins.

protein class (*e.g.* membrane proteins) or set of processes (*e.g.* transcription). The occurrence and distributions of simple sequences in these proteins were then analyzed using the approaches used on intact proteomes above. The results are shown in Table II. Note that the data shown are highly dependent upon the completeness and accuracy of the annotations in SWISS-PROT as well as how well studied the particular classes of proteins are in each organism. As a result of these limitations we have found comparatively few protein sequences in most cases. In addition, some proteins may appear in more than one classification in Table II. Thus, it is difficult to make direct comparisons between classes as to the number of simple sequences found as well as between organisms. However, it is feasible to consider the most common types of simple sequence found (Table II).

Immediately noticeable in Table II is the abundance of serine-rich sequences in almost all classes of proteins examined. Serine-rich sequences are the most common in all four organisms, particularly at the most abundant short length scales (Fig. 4), so perhaps this finding is not surprising. However, the role of serine-rich sequences is not clear. As noted above, it has been proposed that such sequences can act as flexible linkers between protein domains (14) or as protein interaction domains (29, 30). Proteins containing regions enriched in both serine and arginine have also been shown to be involved in mRNA splicing control (37). Serine-rich regions may also function as some form of phosphorylation switch, much as the C-terminal domain of RNA polymerase II operates (29).

Considering now each class of protein in Table II, it can be seen that the most common simple sequences in the limited set of cell cycle proteins identified are serine-rich. Very few cell cycle proteins were found except in the case of SC, which is perhaps the model system for studying these processes. The 102 SC cell cycle proteins found possess a total of 177 simple sequences, over one-third of which (64) are serine-rich. This is a clear enrichment of such sequences as compared with the overall distribution of simple sequences in SC (Fig. 4). Potential roles for these sequences are as discussed above.

We found relatively few proteins with the keyword "metabolism" in their annotations (Table II). With the preceding finding as a caveat, it is notable that there are fewer simple sequences per protein in metabolism-related proteins (significantly less than one per protein) than the average over intact proteomes (slightly more than one per protein; Table I). This would suggest that simple sequences are either generally not required in metabolism-related proteins or that they are discriminated against in comparison to other protein classes. However, as already noted, few proteins were identified in this class, and we could simply be observing the vagaries of poor statistics.

Using "signal" as a keyword, we have identified a significant number of proteins in all four eukaryotes (Table II). These proteins possess a significant number of simple sequences, the most common of which are enriched in serine, threonine, proline, and perhaps surprisingly leucine. Given that signal transduction processes involve significant numbers of phosphorylation and dephosphorylation events, it is perhaps not so remarkable that serine- and threonine-rich sequences are common in signaling proteins. There are also a number of small protein interaction domains common in signaling processes (*e.g.* Src homology 3 domains) that bind to proline-rich sequences (11), leading to an enrichment in such sequences in this class. Thus, the occurrence of serine-, threonine-, and proline-rich sequences in this class of proteins would appear to be biologically significant. The occurrence of a significant number of leucine-rich sequences is at first puzzling, particularly given that such sequences are found at levels lower than would be predicted using our Poisson-distribution model (Fig. 4). However, it is possible that a reasonable number of the proteins in this class possess membrane-spanning segments that, as will be discussed below, can be leucine-rich.

A significant number of transcription-related proteins were also identified (Table II). Remarkably, the transcription-related proteins in SC and DM possess enormous numbers of simple sequences (677 in 274 proteins and 838 in 177 proteins, respectively). Although the same level of enrichment is not seen in CE and AT, it is tempting to postulate that large numbers of simple sequences indicate important functional roles in transcription processes. Indeed, such proteins are known to often possess glutamine-rich sequences (3), so it is not surprising that such sequences are common in DM transcription-related proteins. We also find large numbers of serine-rich sequences (Table II). Perhaps the best known example of a serine-rich region acting as a phosphorylation switch is in RNA polymerase II (29). Although not enriched in serine enough to be found in our surveys, this region is known to interact with a variety of transcription factors when not phosphorylated. These interactions, and consequently transcription, are interrupted when serines become phosphorylated. It is possible that there are similar serine-rich switch/interaction regions in other transcription-related proteins.

A reasonable number of transport-related proteins were also found (Table II). These possess approximately the same numbers of simple sequences as would be expected from the overall average values for the four eukaryotes (Table I). Leucine-, alanine-, and serine-rich sequences are the most common. A significant number of transport-related proteins will be associated with membranes given that transport of molecules through membranes is a common and vital set of processes. The large numbers of leucine-rich and perhaps alanine-rich regions are then most likely indicative of membrane-spanning regions as suggested by Schwartz *et al.* (28).

Finally, we have identified numerous membrane-associated proteins, many of which contain simple sequences (Table II). Presumably for the reasons noted above, large numbers of leucine-rich sequences are found in this class of proteins. In fact, many of these leucine-rich regions are annotated as being membrane-spanning in the SWISS-PROT files for these proteins. Why serine-rich regions would be so abundant is not clear. Some of these are probably found in signaling proteins associated with the membrane (see above), while others may be acting as flexible linkers separating soluble domains from integral membrane domains. Wootton and Drummond (14) have hypothesized that serine-rich regions act as flexible linkers. Notably, glycine-rich regions, also thought to act as linkers, are common in DM membrane-related proteins. Perhaps serine-rich regions are substituted for glycine-rich in the other organisms (Table II).

Simple Sequence Structure—It would of course be useful to know the types of structures adopted by protein simple sequences. Unfortunately little is known about the structural properties of such sequences. Saqi (17) and more recently Huntley and Golding (38) have looked for all occurrences of simple sequences in protein structures in the Protein Data Bank (39). Very few were found. Huntley and Golding (38) point out that simple sequences are under-represented in the Protein Data Bank and hypothesize that this indicates that such regions are intrinsically disordered. Intrinsically disordered regions of proteins are a barrier to structure determination and are consequently routinely deleted from proteins by structural biologists. That simple sequences, particularly relatively long sequences, are disordered is supported by the work of Dunker and co-workers (15, 16, 40), who use low complexity sequences as identifiers of intrinsically disordered proteins. There are indications, however, that not all protein simple sequences are unstructured. For example, leucine-rich membrane-spanning sequences will be highly structured, most likely α-helices, in the membrane. Proline-rich regions are believed, and in many cases have been shown, to adopt the left-handed polyproline II helical conformation (11). It would be a mistake to assume that all simple sequences are unstructured. This is an area that clearly requires further investigation.

CONCLUSIONS

We have presented here a survey of protein simple sequences highly enriched in a single residue type ($\geq$50%) in the proteomes of four eukaryotes. For comparison we have also surveyed the proteomes of 26 prokaryotes. A strikingly large number of simple sequences are found in all of the organisms surveyed (Table I). We find that eukaryotes possess, on average, one or more such simple sequences per protein, whereas prokaryotes average less than one simple sequence for each protein in their proteomes. Furthermore, proteins in eukaryotes that possess at least one simple sequence average between just under two up to slightly more than three simple sequences per protein. These findings are consistent with the work of others (19, 21, 23). The number of simple sequences in the proteomes of prokaryotes is strongly correlated with the number of proteins in their proteomes (Fig. 1*b*). Given that we have only surveyed four proteomes, it is not clear that a linear relationship will be applicable to eukaryotes.

Among the eukaryotes we find that DM possesses more simple sequences per protein than any of the other three eukaryotes (Table I). This is true for all simple sequence lengths (Fig. 3). By comparison, SC, CE, and AT possess a similar number of simple sequences per protein at most lengths with SC perhaps showing some preference for long simple sequences (Fig. 3). In the distributions for the intact proteomes, we find that simple sequences enriched in certain residues, for example alanine, glutamine, glutamate, glycine, and serine, appear to be favored, whereas other residues, specifically leucine, isoleucine, and valine, are discriminated against. These preferences do not correlate with residue occurrence. Some of these observed preferences can be rationalized in terms of structure and/or function, while others remain enigmatic.

The most notable finding of these surveys is that each of the eukaryotes possesses its own unique distribution of protein simple sequences. We find that each organism apparently has preferences for simple sequences enriched in certain residues while at times disfavoring simple sequences enriched in other residues. It is not clear why these eukaryotes have evolved to have differing simple sequence distributions. However, given the sheer number of such sequences found plus the known functional importance of those simple sequences that have been studied in detail, it is tempting to postulate that not only have eukaryotes evolved to tolerate large numbers of simple sequences but also that they require many of these. A simple analysis of simple sequences in classes of proteins indicates that some classes may favor simple sequences enriched in certain residues (Table II).

The data presented here raise questions that can only be answered by further study and analysis. For example, is there an association between type of simple sequence and function? The data in Table II are suggestive but by no means conclusive. Do different organisms use different types of simple sequence for the same function? The fact that each organism possesses a unique distribution implies that this may be the case, but we have no direct evidence. What are the structural properties of such sequences? Little structural data is currently available, although it is clear that it would be incorrect to assume that all simple sequences will be disordered. Answers to questions such as these will shed light on the abundance and distributions of simple sequences highlighted here.

Acknowledgments—We thank Brian Chellgren, Marnie Campbell, David Rodgers, Rajeev Aurora, and George Rose for helpful discussions.

* This work was supported in part by National Science Foundation Grant MCB-00110720 (to T. P. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Supported in part by a postdoctoral research grant in computational genomics from the Pharmacia Corp.

§ To whom correspondence should be addressed: Center for Structural Biology, Dept. of Molecular and Cellular Biochemistry, University of Kentucky, 800 Rose St., Lexington, KY 40536-0298. Tel.: 859-323-6037; Fax: 859-323-1037; E-mail: tpcrea0@uky.edu.

REFERENCES

1. Wootton, J. C., and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266,** 554–571
2. Brendel, V., and Karlin, S. (1989) Association of charge clusters with functional domains of cellular transcription factors. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 5698–5702
3. Gerber, H. P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., and Schaffner, W. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263,** 808–811
4. Kashi, Y., King, D., and Soller, M. (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* **13,** 74–78
5. Katti, M. V., Sami-Subbu, R., Ranjekar, P. K., and Gupta, V. S. (2000) Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci.* **9,** 1203–1209
6. Green, H., and Wang, N. (1994) Codon reiteration and the evolution of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 4298–4302
7. Cummings, C. J., and Zoghbi, H. Y. (2000) Trinucleotide repeats: mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet.* **1,** 281–328
8. Karlin, S., and Burge, C. (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 1560–1565
9. Michelitsch, M. D., and Weissman, J. S. (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 11910–11915
10. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J., and Gentles, A. J. (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 333–338
11. Kay, B. K., Williamson, M. P., and Sudol, M. (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.* **14,** 231–241
12. Williamson, M. P. (1994) The structure and function of proline-rich regions in proteins. *Biochem. J.* **297,** 249–260
13. Karlin, S. (1995) Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.* **5,** 360–371
14. Wootton, J. C., and Drummond, M. H. (1989) The Q-linker: a class of interdomain sequences found in bacterial multidomain regulatory proteins. *Protein Eng.* **2,** 535–543
15. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein. *Proteins* **42,** 38–48
16. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11,** 161–171
17. Saqi, M. (1995) An analysis of structural instances of low complexity segments. *Protein Eng.* **8,** 1069–1073
18. Meyer, E. F., and Tollet, W. J., Jr. (2001) WWWWhy does nature stutter? A survey of strands of repeated amino acids. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **57,** 181–186
19. Huntley, M., and Golding, G. B. (2000) Evolution of simple sequence in proteins. *J. Mol. Evol.* **51,** 131–140
20. Golding, G. B. (1999) Simple sequence is abundant in eukaryotic proteins. *Protein Sci.* **8,** 1358–1361
21. Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1999) A census of protein repeats. *J. Mol. Biol.* **293,** 151–160
22. Katti, M. V., Ranjekar, P. K., and Gupta, V. S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18,** 1161–1167
23. Nishizawa, K., Nishizawa, M., and Kim, K. S. (1999) Tendency for local repetitiveness in amino acid usages in modern proteins. *J. Mol. Biol.* **294,** 937–953
24. Soper, H. E. (1914) Tables of Poisson's exponential limit. *Biometrika* **10,** 25–35
25. Bork, P., and Copley, R. (2001) Filling in the gaps. *Nature* **409,** 818–820
26. Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T. A., Welti, R., Goo, Y. A., Leithauser, B., Keller, K., Cruz, R., Danson, M. J., Hough, D. W., Maddocks, D. G., Jablonski, P. E., Krebs, M. P., Agevine, C. M., Dale, H., Isenbarger, T. A., Peck, R. F., Pohlschroder, M., Spudich, J. L., Jung, K. W., Alam, M., Freitas, T., Hou, S., Daniels, C. J., Dennis, P. P., Omer, A. D., Ebhardt, H., Lowe, T. M., Liang, P., Riley, M., Hood, L., and DasSarma, S. (2000) Genome sequence of Halobacterium species NRC-1. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 12176–12181
27. White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., Moffat, K. S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J. J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K. S., Aravind, L., Daly, M. J., Minton, K. W., Fleischmann, R. D., Ketchum, K. A., Nelson, K. E., Salzberg, S., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577
28. Schwartz, R., Istrail, S., and King, J. (2001) Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci.* **10,** 1023–1031
29. Corden, J. L. (1990) Tails of RNA polymerase II. *Trends Biochem. Sci.* **15,** 383–387

30. Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R. V., Gentile, C., Gebara, M., and Corden, J. L. (1996) The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 6975–6980
31. Aurora, R., Creamer, T. P., Srinivasan, R., and Rose, G. D. (1997) Local interactions in protein folding. Lessons from the α-helix. *J. Biol. Chem.* **272,** 1413–1416
32. Chakrabartty, A., Kortemme, T., and Baldwin, R. L. (1994) Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **3,** 843–852
33. Kreil, D. P., and Kreil, G. (2000) Asparagine repeats are rare in mammalian proteins. *Trends Biochem. Sci.* **25,** 270–271
34. Tonjum, T., Caugant, D. A., Dunham, S. A., and Koomey, M. (1998) Structure and function of repetitive sequence elements associated with a highly polymorphic domain of Neisseria meningitidis PiLQ protein. *Mol. Microbiol.* **29,** 111–124
35. Bairoch, A., and Apweiler, R. (1997) The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *J. Mol. Med.* **75,** 312–316
36. Bairoch, A., and Boeckmann, B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19,** 2247–2249
37. Manley, J. L., and Tacke, R. (1996) SR proteins and splicing control. *Genes Dev.* **10,** 1569–1579
38. Huntley, M. A., and Golding, G. B. (2002) Simple sequences are rare in the Protein Data Bank. *Proteins* **48,** 134–140
39. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542
40. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.* **19,** 26–59