PROTEINS: Structure, Function, and Genetics 35:440–446 (1999)

A Fast Algorithm for Genome-Wide Analysis of Proteins With Repeated Sequences

Matteo Pellegrini, Edward M. Marcotte,* and Todd O. Yeates
Molecular Biology Institute and UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, University of California, Los Angeles, Los Angeles, California

ABSTRACT **We present a fast algorithm to search for repeating fragments within protein sequences. The technique is based on an extension of the Smith-Waterman algorithm that allows the calculation of sub-optimal alignments of a sequence against itself. We are able to estimate the statistical significance of all sub-optimal alignment scores. We also rapidly determine the length of the repeating fragment and the number of times it is found in a sequence. The technique is applied to sequences in the Swissprot database, and to 16 complete genomes. We find that eukaryotic proteins contain more internal repeats than those of prokaryotic and archael organisms. The finding that 18% of yeast sequences and 28% of the known human sequences contain detectable repeats emphasizes the importance of internal duplication in protein evolution. Proteins 1999;35:440–446.** © 1999 Wiley-Liss, Inc.

Key words: sub-optimal alignments; Poisson statistics; tandem repeats; Smith-Waterman algorithm; internal duplications

INTRODUCTION

One of the major goals of protein sequence analysis, and genomics research in general, is to uncover the evolutionary history of the proteins encoded by genomes. Two mechanisms commonly invoked to account for the complexity and diversity of modern proteins are duplications of entire genes leading to paralogous proteins and duplications of fragments within genes leading to proteins containing repeats.

The prevalence of the former event is uncovered by an exhaustive all versus all comparison of proteins in a genome. Such searches reveal that many proteins have paralogues within the same genome. It has been found that the percentage of sequences with paralogues is 46% in *Saccharomyces cerevisiae,* 37% in *Methanococcus jannaschii,* and 30% in *Haemophilus influenzae.*[1] These estimates are necessarily conservative, since other paralogous sequences may have diverged beyond recognition.[2]

Here we present an algorithm to study the second mechanism: the duplication of fragments within a gene. It is known that in higher organisms such events are common.[3] For example, the repeat fragment *GXY*, where *X* and *Y* are preferentially proline and hydroxyproline, is found repeated hundreds of times in collagen proteins. Similarly, the zinc finger domain, approximately 50 amino acids long, is also a very commonly repeated motif found in DNA binding proteins. In this work we attempt to enumerate all the proteins within a genome that contain internal duplications.

As has been demonstrated previously, the existence of internal gene duplications may be uncovered by aligning a protein sequence against itself.[4] This may be efficiently accomplished by using dynamic programming. The output of the alignment may then be analyzed to find the length of the repeating fragment and the number of times it is found in the sequence.

The algorithm we present to perform this task is computationally efficient. The number of steps required scales as N^2, where N is the length of the sequence. This distinguishes it from certain previously published methods, which although more exact, are too computationally intensive to be readily applied to tens of thousands of proteins.[4,5] Furthermore, our method makes no a priori assumptions about the repeat fragment. This sets it apart from other techniques that specifically search databases for previously recognized patterns, such as those typical of coiled coils,[6] or only for short tandem repeats.[7] Finally, unlike most previously published techniques, ours allows us to estimate the statistical significance of the resulting alignments.

We apply this method to search for repeats within the proteins of the 16 complete genomes that were available in August, 1998. We determine what percentage of the proteins contain statistically significant duplications. We then present the distribution of the lengths of the repeating fragments and the number of times they are repeated.

METHODS

Sub-Optimal Alignments

One of the standard methods used to compare two sequences is the Smith-Waterman algorithm.[8] Given an amino acid substitution matrix that provides a score for matching two residues, the algorithm produces the optimal alignment between two fragments of the sequences. This alignment maximizes the total score between the matched residues while allowing the insertion of gaps. This is accomplished in N by M steps, where N and M are

The first two authors contributed equally to this work

*Correspondence to: Edward M. Marcotte, Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095-1570.

Received 5 October 1998; Accepted 28 January 1999

the lengths of the two sequences, by stepping through $H_{i,j}$, a matrix of optimal partial scores up to position i in sequence 1 and j in sequence 2.

When one compares a protein sequence against itself, the optimal alignment is trivially found to be the matching of every residue with itself (represented by the diagonal of matrix $H_{i,j}$). We are not interested in this alignment, but in all other sub-optimal (off-diagonal) alignments where these self-matches are not allowed. If one fragment of a protein, from residue i to i', matches another fragment, from j to j', where $i \neq j$ and $i' \neq j'$, then we say that some portion of this fragment has been duplicated within the sequence.

To uncover all such sub-optimal alignments we use a modified version of the Smith-Waterman algorithm.[9] This algorithm allows us to find the optimal alignment by stepping through $H_{i,j}$ as before. Once this path is found, it is eliminated from further consideration by modifying the matrix $H_{i,j}$ so that the next best path may not intersect it. This is then repeated until all sub-optimal paths above a certain score are found.

The number of steps required to modify $H_{i,j}$ in order to eliminate a previous path is l^2, where l is the length of the path. Typically l is significantly shorter than N, the full length of the sequence. Therefore the new algorithm is nearly as efficient as the original Smith-Waterman method.

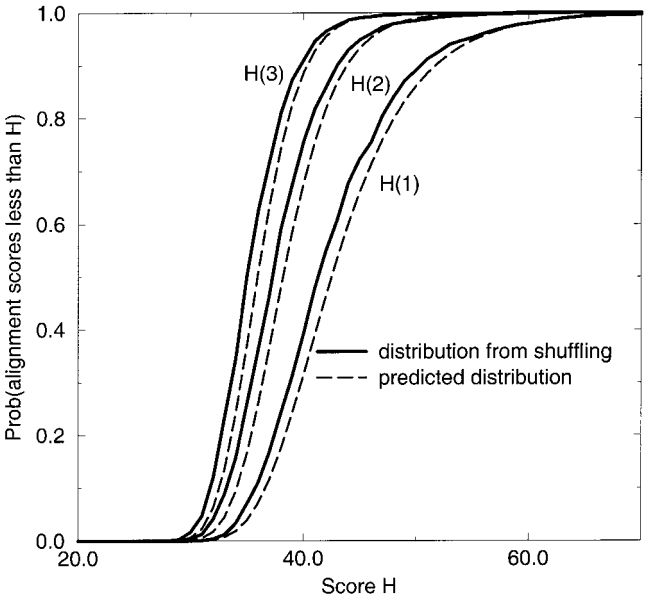

Fig. 1. A comparison of the distributions of scores for the top three sub-optimal alignments (H_1, H_2, and H_3) computed by shuffling the sequence of the human prion protein (thick line) and by applying Eq. (1) (dashed line). The two parameters of the distribution described by Eq. (1) were obtained by fitting the distribution of the first sub-optimal alignments to the first sub-optimal scores of shuffled sequences.

Distribution of Sub-Optimal Scores

In order to compute the statistical significance of a sub-optimal alignment score using the above algorithm, we need to first compute the distribution of scores for random sequences. It is well established that in the case where the average score of a substitution matrix is negative, high scoring alignments are rare and therefore obey Poisson statistics.[10] If the sub-optimal alignments are non-intersecting, as explained above, their scores also obey Poisson statistics. It has been shown by Waterman and Vingron that:

$$P(S_{(k)} < t) = e^{-\gamma N^2 p^t} \sum_{j=0}^{k-1} \frac{(\gamma N^2 p^t)^j}{j!}, \qquad (1)$$

where $S_{(k)}$ is the score of the k th sub-optimal alignment, p and γ are two parameters describing the distribution, and N is the length of the sequence.

To demonstrate the validity of this formula we have generated the first, second, and third sub-optimal alignment scores for 1,000 shufflings of the human prion protein and plotted them on Figure 1. Every time the sequence is shuffled and compared to itself, the diagonal, or self-match is neglected and the next three best alignments are computed using the above algorithm.

In order to compare the computed distributions to the predicted one (Eq (1)) we must estimate the value of the two parameters γ and p. These may be computed by several methods. Generally, the most reliable estimates are obtained by fitting the distribution of optimal scores to the theoretical distribution. We use the method of moments to compute the value of the parameters from the mean and variance of the distribution,[11] and plot the resulting curves in Figure 1. From the figure we see that the theoretical distributions match the computed ones well, and that the estimations of significance levels improve for higher scoring alignments.

To obtain reliable estimates of the two parameters γ and p by the above method we need to shuffle the sequences several hundred times, thus making the overall alignment algorithm correspondingly slower. We use a faster method for parameter estimation in genome searches to speed up the calculations. This method relies on the calculation of the distribution of 1,000 sub-alignment scores from only a few shufflings of the sequence of interest. The average number of scores above a threshold t is given by the mean of the Poisson distribution $\gamma N^2 p^t$.[10] We therefore compute the log of the number of occurrences of a score, make a linear fit and estimate the two parameters by linear regression. In previous work[12] we have shown that the fractional error,

$$E = \sum \frac{|\log_{10}(P_{slow}) - \log_{10}(P_{fast})|}{|\log_{10}(P_{slow})|}, \qquad (2)$$

between the probability estimates obtained by computing the parameters using these two methods is 0.13.

Calculation of Repeat Unit

For sequences which contain repeats, we would like to know the length of the repeating unit and the number of times it is found in the sequence. Many methods have been used to obtain this information in great detail (e.g. ref. 4).

Here we will present two simple methods that allow us to estimate the repeat length and number with minimal computation.

In real proteins, evolution leads to differences between the repeating segments and even to variation in the spacing between them. As a consequence, it is sometimes difficult to reconstruct automatically the exact repeating structure of the protein. We have found empirically that we obtain better estimates of the repeating unit when we analyze a high scoring subset of the sub-alignments. This subset is generated by using only the sub-alignments whose scores are statistically significant as optimal alignments. For example, the human prion protein has about one hundred statistically significant sub-alignments. However only five of these have scores that would be significant optimal alignment scores ($k = 1$ in Eq. (1)).

The first method for extracting the repeat length relies on the analysis of the path matrix, $P_{i,j}$. This $N \times N$ matrix has a value of 1 at the (i, j) position if one of the statistically significant paths passes through this point and has a value of zero otherwise. We sum the autocorrelation function in each column,

$$AC(l) = \sum_{i=1}^{N} \sum_{j=1}^{N-l} P_{i,j} P_{i,j+l}, \tag{3}$$

and find the repeat length that has the highest peak. This method is able to accurately assess the separation between corresponding points in the repeating units of proteins containing tandem repeats. However, if the repeats are interrupted by non-repeating sequence segments, the separation between the repeating units may be an overestimate of the actual length of the conserved repeat.

To address this problem, we have devised a second method that considers more directly the length of conserved repeats. We create a vector of dimension equal to the sequence length. Each time a sequence position is found in a sub-alignment, the corresponding position in our vector is incremented. This may be seen as a projection of the path matrix, $P_{i,j}$, into one dimension:

$$S_i = \sum_{j} P_{i,j} \tag{4}$$

In a typical case of a sequence with multiple repeats, the final entries in our vector range from 0 to approximately twice the number of sub-alignments. When the value of the ith vector position is plotted versus i, the resulting graph resembles a step function. This function has been plotted in Figure 2 for three sequences that contain two, eight and hundreds of repeats.

In an ideal case where each repeating unit in a set of consecutive repeats aligns to another with a statistically significant score, the step function would be flat, since each repeat would participate the same number of times in an alignment. In cases with real proteins, the peak arises because paths which overlap single repeats do not usually produce significant alignments, and therefore the middle

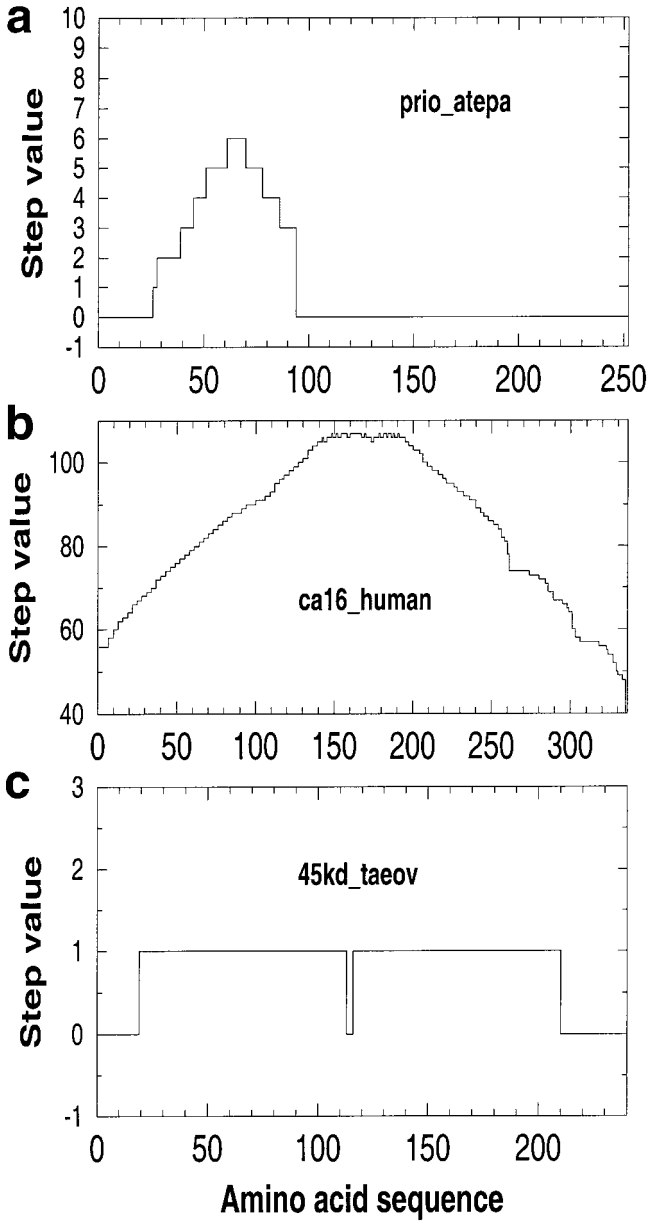

Fig. 2. Examples of step functions generated by projecting the sub-optimal alignments into one dimension. The three proteins are: **(a)** spider monkey prion, **(b)** human collagen, and **(c)** 45 kD antigen from sheep tapeworm. The function is computed by overlapping all sub-optimal alignments with high scores and counting the number of occurrences of each amino acid position. The resulting repeat lengths are 8, 3, and 95.

repeats are found more often in the sub-optimal paths. Also, real proteins often contain non-repeating sequences between repeats which lead to a separation between steps.

From this function we compute the average length of a step, which corresponds to the length of the repeat unit. The steps are sequential positions in the vector with equal values of the counter, which must be greater than zero. The average length of the step is computed. For cases in which the repeat unit was known, we found that a better estimate of the true repeat length was found by excluding

steps lengths that are further than two standard deviations from the mean.

When we compare the two methods, we expect the autocorrelation analysis to be more accurate for short tandem repeats. Conversely, as the repeating units grow in length and are more likely to be separated by non-repeating fragments, we expect the step function method to yield better estimates. In the results that follow, we apply the step function approach to compute the distribution of repeat lengths in complete genomes, and use the autocorrelation method to compute in more detail the distribution of short (<20 amino acid) repeats.

Two automatic methods were investigated for obtaining the number of repeats, n, in a sequence. In the first, we make use of the expectation that the sum of all the path lengths should be approximately $n^2 l$,

$$n \approx \left(\frac{\sum_i \sum_j P_{i,j}}{l} \right)^{1/2}, \qquad (5)$$

where l is the repeat length. In the second method, we approximate n as the length of the non-zero path in the step function divided by l. The two methods produce similar results. Therefore in the results that follow, the analysis of repeat lengths and number of repeating units was conducted using the step function method.

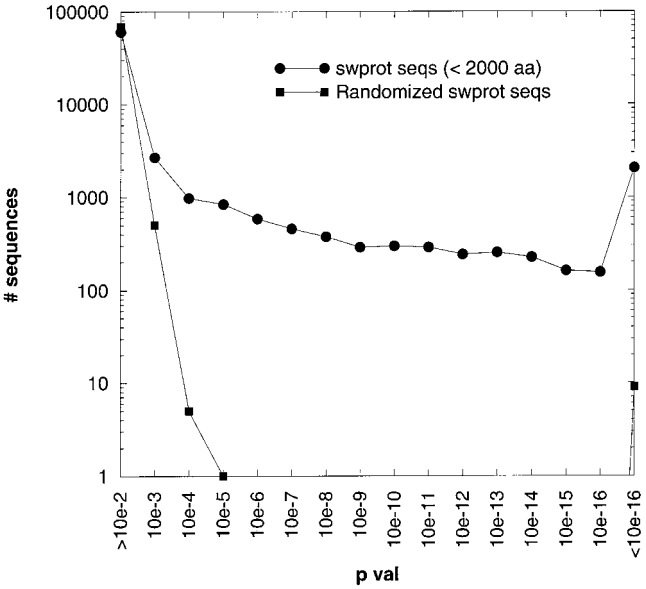

Fig. 3. A histogram of the number of sequences whose first sub-optimal alignments had a given probability value. The curve with dots is computed using the real Swissprot database and the curve with squares is computed for shuffled versions of Swissprot proteins. The area between the two curves gives an estimate of the total number of sequences with repeats.

RESULTS

We first searched for repeats within the Swissprot database. Release 35 of this database contains 71,248 protein sequence entries. To speed up the calculations, we eliminated all sequences longer than 2,000 amino acids, leaving 70,822 proteins. For each of these we compute the statistical significance (P value) of the first sub-optimal alignment. The histogram of these values on a $log - log$ plot is shown in Figure 3.

To ensure the validity of our P value calculation we repeated this same procedure on shuffled Swissprot sequences. The histogram of these result is also shown on Figure 3. For P values of 10^{-3} we have an approximately tenfold excess in the real Swissprot versus the shuffled sequences. This implies that sequences with this P value have a 10% false positive rate. For future calculations we set this value as a threshold.

To estimate the total number of sequences with detectable repeats within Swissprot, we subtract the histogram of shuffled sequences from that of real sequences. This accounts for the rate of false positives. This calculation leads to an estimate of 9,926 sequences with repeats, or 14% of the entire database. This number represents a conservative estimate of repeat percentage since evolution may have obscured repeats within other proteins. The algorithm recovers 77% of the 4,351 sequences annotated within Swissprot as containing repeats, missing cases where repeats are known only from protein structure or are statistically insignificant. We also find over 6,500 sequences with repeats that are not annotated as containing repeats. Within Swissprot, all these repetitive fragments account for 7% of all amino acids. These observations suggest that internal gene duplication is a common mechanism in the evolution of proteins.

We find a few shuffled sequences that have extremely low P values. These sequences are unusually short, about ten residues. Because the corresponding $H_{i,j}$ matrix contains so few entries, estimates of the P values are inaccurate. Fortunately, less than ten sequences out of the entire Swissprot database were too short for our method to yield reliable statistical estimates of the alignment significance.

We next searched for repeats within the open reading frames of fully sequenced genomes that were available in 1998. This set includes 16 genomes of archael, prokaryotic, and eukaryotic organisms (Genome sequences were obtained from the Institute for Genomic Research and from Fitz-Gibbon[13]). The results are reported in Figure 4. In this plot we also report the results from the search of all of the Swissprot sequences and the subset of Swissprot sequences from humans.

For different organisms, the percentage of sequences containing repeats ranges from 6% to 28%. These percentages are smallest for the archael and prokaryotic organisms, and dramatically higher for the subset of Swissprot proteins from humans. Although this subset contains only a small fraction of the entire human genome, the 28% estimate is probably a reasonable one for the total fraction of sequences with repeats, since there is no reason to assume that this set is systematically biased in favor of repeat-containing proteins. The yeast genome has been fully sequenced and 18% of its sequences contain repeats. The human and yeast results suggest that eukaryotes in

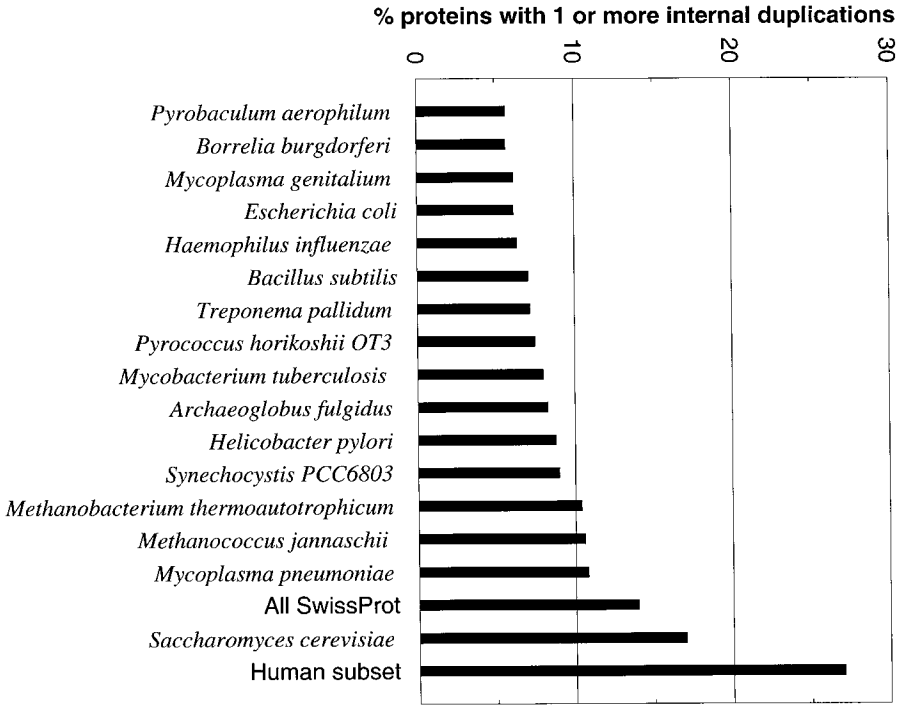

Fig. 4. The percentage of proteins within each genome that contain at least one statistically significant internal duplication. We also report the percentage for all of Swissprot and for the subset of human proteins within Swissprot.

general may contain a larger fraction of proteins with repeats compared to archael and prokaryotic organisms.

We have conducted a more in-depth analysis of the length and number of repeat units within the genomes of *S. cerevisiae, E. coli,* and *A. fulgidus.* These three organisms were selected to obtain one representative from eukaryotes, prokaryotes, and archae. The mean length of the repeating unit and the number of times it is found within a sequence is obtained by the analysis of the step function described in Methods.

In Figure 5(a) and (b) we plot the resulting distributions of the length of the repeating unit for the three genomes. The histogram of Figure 5(a) plotted on a log-linear scale yields a linear trend, implying that the probability of obtaining longer repeating units decays exponentially with length. This trend is particularly noticeable in the *S. cerevisiae* genome, but holds approximately for the other two as well. It is likely that short and long repeats are generated by different mechanisms: short repeats may be due to stutters during replication while longer repeats are more likely to be produced by recombination. It is surprising, therefore, to observe the linear trend of Figure 5(a) over a broad range of repeat lengths.

In Figure 5(b) we notice that the eukaryotic genome has a higher percentage of short (1–3 amino acid) repeat lengths than the prokaryotic and archael genomes. These yeast proteins contain low complexity regions, such as single amino acid stutters, that apparently are uncommon in the other two genomes.

The histogram of the number of repeating units in a sequence is shown in Figure 6. As in Figure 5(a), the data are plotted on a log-linear scale. As a consequence of the eukaryotic genome having a higher percentage of short repeating units than the prokaryotic and archael ones, we find that *S. cerevisiae* also has on average more repeating units in its proteins than *E. coli* and *A. fulgidus.* This fact may also follow from the observation that *S. cerevisiae* proteins are on average longer than those found in *E. coli* and *A. fulgidus.*

CONCLUSIONS

We have developed a fast algorithm to search for internally repeated sequences within large protein databases. The algorithm searches for all statistically significant sub-optimal paths which correspond to internally repeated protein fragments. We are able to assign probability values to each sub-optimal alignment based on Poisson statistics. We are then able to estimate the length of the repeating unit and the number of repeats in the protein sequence.

We have applied these techniques to first evaluate the abundance of repeating sequences within the Swissprot database. By comparing the P values of sub-optimal alignments found with real versus shuffled sequences we are able to confidently assign a P value threshold. Based on these statistical arguments we find that 14% of the Swissprot sequences contain detectable internal repeats.

Next we searched within the open reading frames of 16 complete genomes for proteins with internal repeats. We

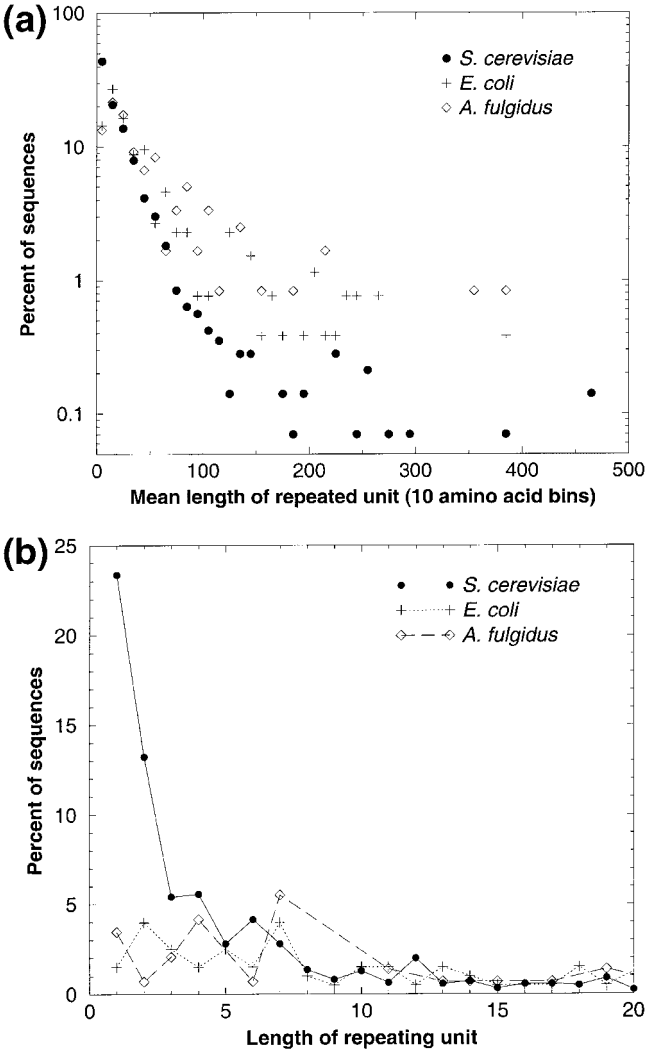

Fig. 5. The percentage of proteins within the genomes of *S. cerevisiae, E. coli,* and *A. fulgidus* as a function of the length of the repeating unit. In **(a)** we use the step function method to compute the lengths, while in **(b)** we focus on the distribution of short repeat lengths using the autocorrelation approach.

find that these percentages range from 6% to 28%. The *S. cerevisiae* genome and especially the subset of human proteins within Swissprot contain a significantly higher percentage of proteins with repeats than the genomes of prokaryotic and archael organisms. Although the number of complete eukaryotic genomes is still limited, this result suggests that eukaryotic proteins are far more likely to contain repeats than prokaryotic or archael ones.

It is interesting to note that the percentage of proteins with internal repeats is smaller than the percentage of proteins that have arisen by duplication of entire genes. For *S. cerevisiae,* these numbers are 18% versus 46%.[1] It may be that entire gene duplications are more easily detected than internal repeats, since on average they involve the duplication of longer sequences. Another possibility is that entire gene duplications are on average less likely to be selected against than internal ones. This could

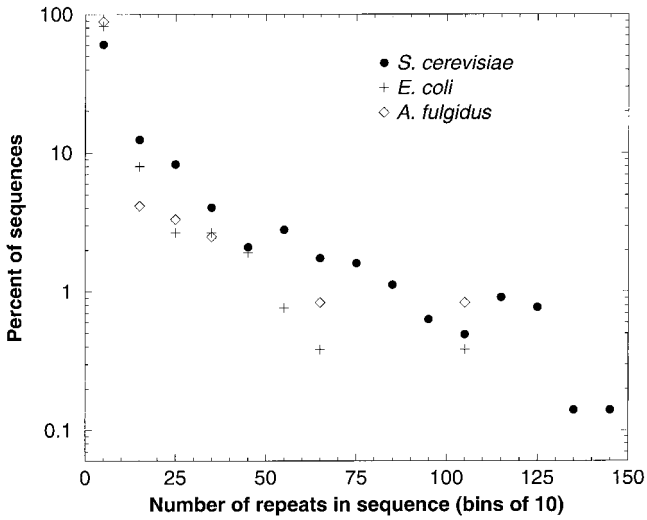

Fig. 6. The percentage of proteins within the genomes of *S. cerevisiae, E. coli,* and *A. fulgidus* as a function of the number of times the repeating unit is found in the protein sequence.

be explained by the tendency of internal duplications to disrupt a protein's structure or function. For instance, in special cases, excessive internal duplication can lead to aggregation.[14]

Finally we analyzed three complete genomes in more detail, to extract the distribution of repeat lengths and number of repeats within the sequences. This analysis reveals that the proteins of *S. cerevisiae* are more likely to contain short (<4 amino acid) repeating units than those of *E. coli* and *A. fulgidus.* Consequently, the eukaryotic proteins seemed to contain a greater number of repeating units than the prokaryotic and archael ones.

The analysis we have conducted begins to reveal the abundance of internally duplicated fragments within protein sequences. The evidence suggests that these events are extremely common, especially in eukaryotic organisms. As more complete genome sequences become available, we hope to further analyze the role of internal duplications in the evolution of modern organisms, possibly incorporating multiple sequence alignments into the analysis. (The program has been made available at the following URL: http://www.doe-mbi.ucla.edu/people/matteo//repeats.html.)

ACKNOWLEDGMENTS

This work is supported by a Sloan Foundation, Department of Energy (DOE) postdoctoral fellowship (M.P.), by a DOE, Oak Ridge Institute for Science and Education Hollaender postdoctoral fellowship (E.M.), and by USPHS grant GM31299 (T.O.Y).

REFERENCES

1. Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archael genomes in terms of protein structure. J Mol Biol 1997;274:562–576.

2. Brenner SE, Hubbard T, Murzin A, Chothia C. Gene duplications in H. influenzae. Nature 1995;378:140.
3. Heringa J, Taylor W. Three-dimensional domain duplication, swapping and stealing. Curr Opin Struct Biol 1997;7:416–421.
4. Heringa J, Argos P. A method to recognize distant repeats in protein sequences. Proteins 1993;17:391–411.
5. Benson G. Sequence alignment with tandem duplication. J Computat Biol 1997;4:351–367.
6. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.
7. Claverie JM, States DJ. Information enhancement methods for large scale sequence analysis. Comput Chem 1993;17:191–201.
8. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
9. Waterman MS, Eggert M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. J Mol Biol 1987;197:723–728.
10. Waterman MS, Vingron M. Sequence comparison significance and Poisson approximation. Stat Sci 1994;9:367–381.
11. Altschul SF, Erickson BW. A nonlinear measure of subalignment similarity and its significance levels. Bull Math Biol 1986;48:617–632.
12. Pellegrini M, Yeates TO. Searching for distant evolutionary relationships between protein sequence families. In press.
13. Fitz-Gibbon ST. Genome sequence of the hyperthermophilic archaeon pyrobaculum aerophilum. Doctoral thesis, University of California, Los Angeles, 1998.
14. Wells RD. Molecular basis of genetic instability of triplet repeats. J Biol Chem 1996;271:2875–2878.