OMICS A Journal of Integrative Biology
Volume 6, Number 1, 2002
Mary Ann Liebert, Inc.

Spectral Analysis of Distributions: Finding Periodic Components in Eukaryotic Enzyme Length Data

EUGENE KOLKER,[1] BRIAN C. TJADEN,[2,1] ROBERT HUBLEY,[1] EDWARD N. TRIFONOV,[3] and ANDREW F. SIEGEL[4,1]

ABSTRACT

We introduce the spectral analysis of distributions (SAD), a method for detecting and evaluating possible periodicity in experimental data distributions (histograms) of arbitrary shape. SAD determines whether a given empirical distribution contains a periodic component. We also propose a system of probabilistic mixture distributions to model a histogram consisting of a smooth background together with peaks at periodic intervals, with each peak corresponding to a fixed number of subunits added together. This mixture distribution model allows us to estimate the parameters of the data and to test the statistical significance of the estimated peaks. The analysis is applied to the length distribution of eukaryotic enzymes.

INTRODUCTION

THE TRADITIONAL APPROACH to extracting and evaluating periodical patterns is harmonic Fourier analysis, whereby the interval over which the experimentally observed function is defined is considered as one period of the extended repeating pattern. The theorem of Fourier holds that any such pattern can be decomposed into a number of harmonics, sine and/or cosine waves, consisting of integer multiples of the basic frequency, with estimated amplitudes and phases. The spectrum derived is, thus, discrete, and the interval under study contains integral numbers of the periods of the harmonic components. Experimental reality may offer empirical distributions (histograms) with periodic components where the experimental interval does not necessarily span an integral number of relevant periods. In this case, the discrete Fourier transform provides only nearest discrete estimates of the noninteger period.

An exact noninteger value (rather than an integer estimate) can be estimated by fitting sine and/or cosine waves with a continuously adjustable period to the experimental data. The fitted values for the tested periods would form a virtually continuous spectrum of the experimental distribution. This is the basis of the spectral analysis of distributions (SAD) technique to detect and evaluate periodic signals in empirical distributions. For each period under consideration, SAD estimates its strength in the observed data.

As a case study to illustrate the applicability of SAD, we have chosen the protein size distribution for eukaryotic enzymes. In earlier work (Berman et al., 1994), the distributions derived on the basis of data avail-

[1]The Institute for Systems Biology, Seattle, Washington.
[2]Computer Science and Engineering, University of Washington, Seattle, Washington.
[3]Department of Structural Biology, The Weizmann Institute of Science, Rehovot, Israel.
[4]Departments of Management Science, Finance, Statistics, and Genome Sciences, University of Washington, Seattle, Washington.

able at that time showed that oscillation was present with a period of approximately 120–125 residues for eukaryotic proteins, in accordance with speculation by Svedberg (1929). These observations are addressed again in this study, with a substantially larger ensemble of sequences. In addition, to avoid possible biases due to a small number of overrepresented protein families, a nonredundant data set of eukaryotic enzymes was constructed by removing all protein entries with the same description, leaving only one representative. Clearly, when the data sample size is larger, the task of detecting and then evaluating whether there are any periodic signals in the data becomes less difficult. Of all the enzyme data sets of the different kingdoms, only the eukaryotes have a large nonredundant subset. The possible periodic signal (oscillation) has been approximated by the cosine function. The choice of the cosine here is due to the nature of the expected signal, that is, possible typical protein (domain) unit size and its multiples, with expected positive peaks at multiples of the (possible) typical unit length. After the amplitude of the oscillation is estimated, the question of statistical significance of the observation is addressed using the method of maximum likelihood to evaluate a periodic probability mixture model and to test its significance using a likelihood ratio test.

METHODS AND DATA

SAD algorithm

The distribution of the raw data is represented by the number of occurrences, $Total_i$, of the value i in the data set defined for a relevant interval from $i_{\min}$ to $i_{\max}$. For example, consider the length distribution of eukaryotic enzymes, for which $Total_i$ is the number of enzymes containing i amino acids (Fig. 1). Since the majority of protein lengths are not shorter than 50 amino acids (aa) and not longer than 600 aa (over 75% for both eukaryotic enzyme data sets; see Table 1), we have $i_{\min} = 50$ and $i_{\max} = 600$.

For each period tested, $Total_i$ is modeled as the sum of two parts: (1) the nonoscillating background, $Nonosc_i$, and (2) the oscillating component, Osc_i. First, the nonoscillating part of the original distribution is calculated by smoothing the total using an equally weighted moving average over a sliding window equal in size to the period and centered at i. This window is chosen in order to eliminate any component that is periodic at the period being tested. Thus, for a given period j (which varies in this case from 2 to 200 aa by increments of 1 aa), $Nonosc_i$ is defined within a reduced interval (a complete interval without half-periods from both ends) $[i_{\min} + \text{int}(j/2),\ i_{\max} - \text{int}(j/2)]$:

$$Nonosc_{ij} = \frac{1}{j}\left\{\sum_{k=-\text{int}(j/2)}^{\text{int}(j/2)} Total_{i+k} - \left[\frac{j+1}{2} - \text{int}\left(\frac{j+1}{2}\right)\right]\left[Total_{i-\text{int}(j/2)} + Total_{i+\text{int}(j/2)}\right]\right\} \quad (1)$$

where use of $\text{int}[(j + 1)/2]$ allows us to unify the cases of even and odd period size j.

The smoothing to produce $Nonosc_{ij}$ eliminates all oscillations with period j from $Total_i$. Next, the nonoscillating background is subtracted from the actual distribution to obtain the oscillating component Osc_{ij}:

$$Osc_{ij} = Total_i - Nonosc_{ij} \quad (2)$$

which is defined within the shorter interval of $[i_{\min} + \text{int}(j/2),\ i_{\min} + \text{int}(j/2) + jm]$ where $m = \lfloor (i_{\max} - i_{\min})/j - 1 \rfloor$, the maximal integral number of (full) periods of j within the interval $[i_{\min} + \text{int}(j/2),\ i_{\max} - \text{int}(j/2)]$.

Finally, the resulting oscillating component was subjected to the following cos-Fourier transform that is represented as a cosine function with period j and a remainder part $Rest_{ij}$:

$$Osc_{ij} = A_j \cos(2\pi i/j) + Rest_{ij} \quad (3)$$

where the resulting amplitude A_j can be calculated as a classic Fourier coefficient:

$$A_j = \frac{\sum_{i=i_{\min}+\text{int}(j/2)}^{i_{\min}+\text{int}(j/2)+jm} Osc_{ij} \cos(2\pi i/j)}{\sum_{i=i_{\min}+\text{int}(j/2)}^{i_{\min}+\text{int}(j/2)+jm} \cos^2(2\pi i/j)} \quad (4)$$

where j is a given period value, m is the number of multiples of the period j, and i is the variable coordinate (e.g., the protein length measured in amino acid residues). By definition of the amplitude coefficient

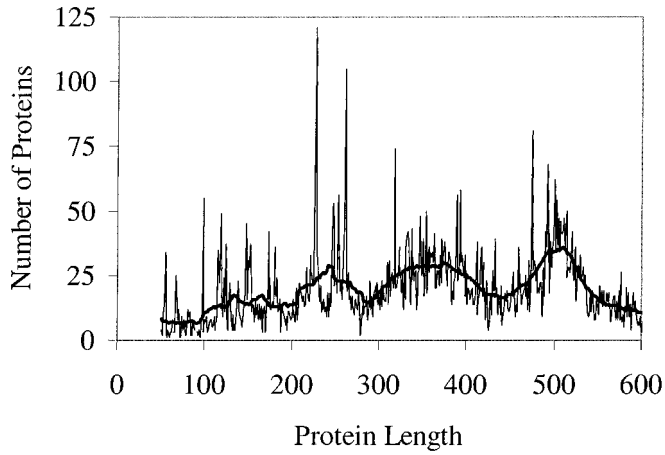

FIG. 1. Distribution of eukaryotic enzyme lengths, entire data set. The histogram corresponds to the actual (raw) distribution, and the smoothed curve corresponds to the running averages with a 41-aa window.

A_j, the component $Rest_{ij}$ does not contribute to the cosine expansion; in other words, this component is orthogonal to these periodic functions.

Assumptions and uncertainties

The SAD approach requires some underlying assumptions regarding the data source and its uncertainty, as is typical of statistical methods:

1. The original data reflect the (biological) mechanism, which might generate an oscillating component at a certain frequency.
2. Along with the given periodic component, the data include some experimental background caused by tendencies of a different nature with no systematic biases.
3. All experimental observations (e.g., protein lengths) are independent and sampled from the same distribution.
4. The total number of data points is large enough to provide a representative sampling of the random uncertainties.
5. According to the sampling theorem (Pohlmann, 2000), more than two data points must exist per period of a resolvable spectral component.

One important feature of SAD is its ability to yield a virtually continuous spectrum rather than discrete harmonics. The virtually continuous spectrum means that steps between periods of interest need not be integers.

Statistical model

We model the probability distribution of the length of a protein chain selected at random from the data set as a statistical mixture of a smooth background distribution together with k individual peaks at the preferred length μ and its multiples 2μ, 3μ, . . . $k\mu$. The strengths of the peaks are p_1, p_2, . . . p_k (expressed as nonnegative probabilities with sum less than 1) and the standard deviations are σ, $\sqrt{2}\cdot\sigma$, . . . $\sqrt{k}\cdot\sigma$ respectively. Intuitively, the model says that, in addition to a basic background distribution, with probability p_i a protein chain is composed of i subunits, each with mean length μ and standard deviation σ and having length statistically independent of the other subunits. The model is illustrated in Figure 2.

In order to adapt to a variety of potential background distribution shapes of varying skewness, we use the gamma distribution family, which has density $x^a e^{-x/\beta}/[\Gamma(\alpha + 1)\beta^{\alpha+1}]$ defined by a shape parameter α and a scale parameter β. Equivalently, the gamma distribution can be parameterized by its mean value $\mu_{background} = \beta(\alpha + 1)$ and standard deviation $\sigma_{background} = \beta\sqrt{\alpha+1}$. We use normal distributions for the peaks

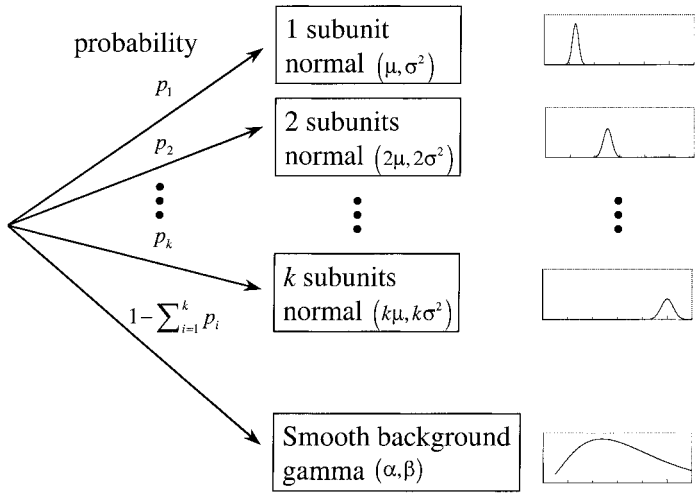

FIG. 2. Mixture distribution model. The model consists of a smooth background together with superimposed peaks in the distribution.

with parameters μ and σ as specified above. In order to convert the continuous distributions of the model to the discrete distributions observed in the data over the range from $i_{\min}$ to $i_{\max}$, we normalize each density function (gamma or normal) by dividing its value at each such positive integer by the sum of all such density values for this range. We denote the resulting normalized probability density functions over integers x from $i_{\min}$ to $i_{\max}$ as $g(x; \alpha,\beta)$ for the gamma distribution, and $f_i(x; \mu,\sigma)$ for the ith normal distribution (normalizing a normal distribution with mean $i\mu$ and standard deviation $\sqrt{i}\cdot\sigma$).

We use the method of maximum likelihood to estimate the $(k + 4)$ parameters μ, σ, α, β, $p_1, \ldots, p_k$. That is, we find the parameter combination most likely to have produced the observed data set. Implementation of this approach to the analysis of eukaryotic enzyme length data is illustrated in Figure 3.

We use a generalized likelihood ratio test to determine significance of the estimated peaks (Kendall and Stuart, 1979). The maximized likelihood for the n data values $x_1, \ldots, x_n$ is

$$L_1 = \max_{\mu,\sigma,\alpha,\beta,p_1,\ldots,pk} \prod_{j=1}^{n} \left[\left(1 - \sum_{i=1}^{k} p_i\right) g(x_j; \alpha,\beta) + \sum_{i=1}^{k} p_i f_i(x_j; \mu, \sigma)\right]$$

corresponding to the unconstrained maximum likelihood estimators $\hat{\mu}$, $\hat{\sigma}$, $\hat{\alpha}_1$, $\hat{\beta}_1$, $\hat{p}_1, \ldots, \hat{p}_k$. Estimated parameters for the gamma distribution $(\hat{\alpha}, \hat{\beta})$ define the background estimates $(\hat{\mu}_{background}, \hat{\sigma}_{background})$. The

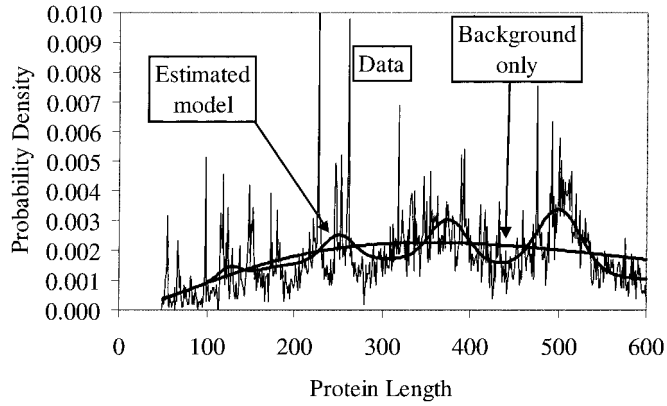

FIG. 3. Estimated probability density of eukaryotic enzyme lengths, entire data set. The data are shown with the estimated model and the constrained model consisting of background only without any peaks.

TABLE 1. NUMBER OF EUKARYOTIC ENZYMES AND THEIR LENGTHS

	Total	*Nonredundant*
Total no. of enzymes	13,613	7,232
≤600 aa	10,741 (78.9%)	5,480 (75.8%)

maximized likelihood L_1 is then compared to the maximum constrained likelihood (pure background, without the peaks) defined as

$$L_0 = \max_{\alpha,\beta} \prod_{j=1}^{n} g(x_j; \alpha, \beta)$$

corresponding to the constrained maximum likelihood estimators $\hat{\alpha}_0$, $\hat{\beta}_0$ (Fig. 3) defining the pure background estimates $\hat{\mu}_{pure\ background}$ and $\hat{\sigma}_{pure\ background}$. The p value is based on the likelihood ratio statistic

$$l = -2\ln(L_0/L_1)$$

which, asymptotically, has a χ^2 distribution with $(k + 2)$ degrees of freedom under the null hypothesis that the chain lengths come from a gamma distribution without any superimposed peaks.

Data

The raw data used in this study was obtained from SWISS-PROT (Baivoch and Apweiler, 2000) Release 39.16 of 12-Apr-2000 (94,743 protein sequences). The eukaryotic sequences were separated from the entire dataset using the SWISS-PROT "OC" database field. Protein fragments were removed by filtering on the word "FRAGMENT" and its derivates in the "DE" database field. Next entries were removed for which no known enzymatic reaction has been identified. This was accomplished by searching for entries which contain Enzyme Commission (EC) numbers in the SWISS-PROT "DE" database field. All remaining enzyme sequences (polypeptide chains) where then counted by length. In addition, to avoid possible biases due to a small number of overrepresented protein families, a nonredundant data set of eukaryotic enzymes was constructed by removing all protein entries with the same description, leaving only one representative with the longest length. Similar results were obtained with either random or shortest length choice of the representative enzyme. Of all the enzyme data sets of the different kingdoms, only the eukaryotes have a large nonredundant subset. A summary of the resulting sizes for eukaryotic enzymes for two data sets can be found in Table 1. The maximal length among eukaryotic enzymes was found to be 5,217 aa, but a majority of them (over 75% for both entire and nonredundant sets; see Table 1) are not longer than 600 aa. Because a majority of entries shorter than 50 aa do not represent full-length proteins, these were not considered. Similar results were obtained when entries with lengths shorter than 60, 70, and 80 aa were re-

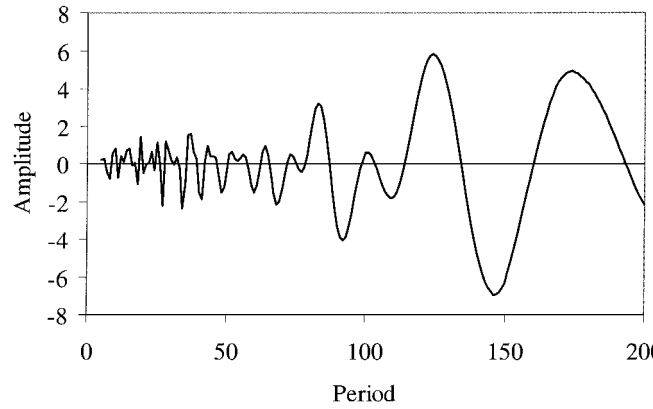

FIG. 4. Cosine spectrum of eukaryotic enzyme lengths, entire data set. The cosine spectrum generated by SAD achieves a maximum amplitude at 124 aa.

TABLE 2. STATISTICAL PARAMETERS AND *p* VALUES

	Total	*Nonredundant*
$\mu_{pure\ background}$	560.14	566.53
$\sigma_{pure\ background}$	371.44	314.57
$\mu_{background}$	479.77	488.32
$\sigma_{background}$	297.95	263.09
μ	125.02	126.80
σ	12.20	8.09
p_1	0.0106	0.0155
p_2	0.0361	0.0175
p_3	0.0743	0.0551
p_4	0.1260	0.1093
p value	2.08×10^{-144}	1.71×10^{-91}

moved. As a result, two data sets of eukaryotic enzymes: entire and nonredundant were subjected to SAD analysis.

RESULTS

Detecting possible periodicities in the lengths of eukaryotic enzymes

The spectral analysis methodology described above begins by detecting a preferred periodic signal component in a raw data histogram. Figure 1 shows the histogram of the size distribution of eukaryotic enzymes. Several maxima (approximately at 125 aa and its multiples) are seen in the smoothed version of the distribution, in apparent periodic succession. These maxima are detected by the SAD procedure. In the cosine spectrum shown in Figure 4, several maxima are seen, of which the one at 124 aa has the highest amplitude. Other maxima have lower amplitudes and will be considered as secondary. The origins of these secondary maxima are not clear and may deserve further study. The appearance of the major maximum in the spectrum suggests a possible periodicity of approximately 125 aa in the lengths of eukaryotic enzymes. Statistical significance of the observed period is determined using likelihood methods as defined for the statistical model.

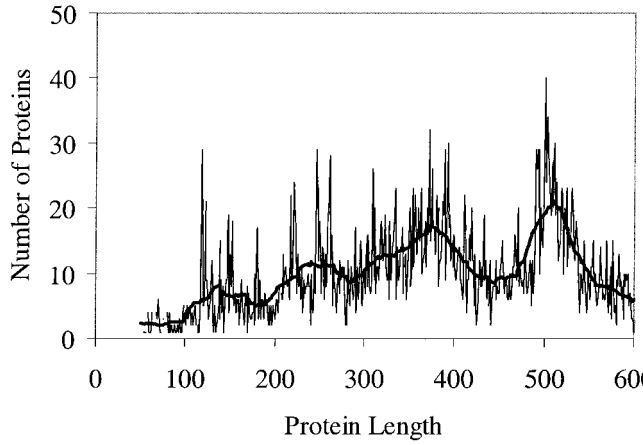

FIG. 5. Distribution of eukaryotic enzyme lengths, nonredundant data set. The histogram corresponds to the actual (raw) distribution, and the smoothed curve corresponds to the running averages with a 41-aa window.

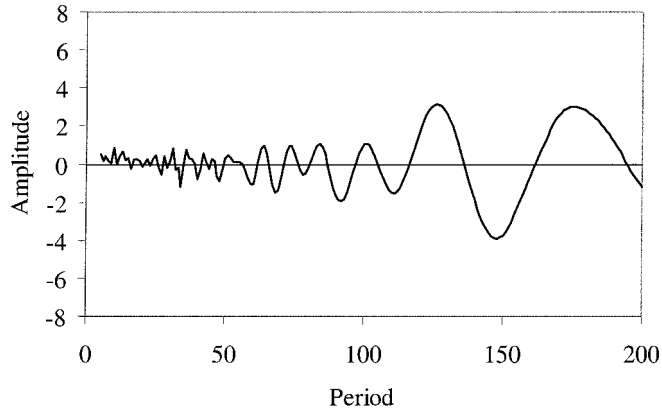

FIG. 6. Cosine spectrum of eukaryotic enzyme lengths, non-redundant data set. The cosine spectrum generated by SAD achieves a maximum amplitude at 126 aa.

Evaluating the statistical significance of the 125-aa period

The statistical model was used to find maximum likelihood estimates for the probability mixture model most likely to have produced the data set. The estimated parameters are summarized in Table 2. These parameters include $\hat{\alpha} = 1.593$ and $\hat{\beta} = 185.031$ for the background gamma distribution; $\hat{\mu} = 125.02$ and $\hat{\sigma} = 12.20$ for the normal distribution of a single subunit length; and subunit strengths $\hat{p}_1 = 0.0106$, $\hat{p}_2 = 0.0361$, $\hat{p}_3 = 0.0743$, and $\hat{p}_4 = 0.1260$. The pure background gamma distribution was estimated as $\hat{\alpha}_0 = 1.69$ and $\hat{\beta}_0 = 208.133$. The estimated model, original data, and pure background are shown in Figure 3. The peaks are very highly statistically significant ($p = 2 \times 10^{-144}$, $\chi^2_6 = 683.64$) based on a likelihood ratio test (Table 2). Such statistical significance provides overwhelming evidence of periodicity, equivalent to a Z statistic of 25.6 (for comparison, a p value of 0.05 corresponds to a Z statistic of 1.96).

Analysis of the nonredundant set

To avoid possible biases among the eukaryotic enzymes, a nonredundant data set was constructed. All proteins with the same description were removed, leaving only one representative enzyme of the longest length. SAD was then applied to the analysis of the non-redundant set of eukaryotic enzymes. Figure 5 shows the raw data histogram and its smoothed distribution. The same four maxima earlier observed in Figure 1 are seen here as well. The cosine spectrum of the eukaryotic enzyme lengths of the nonredundant set

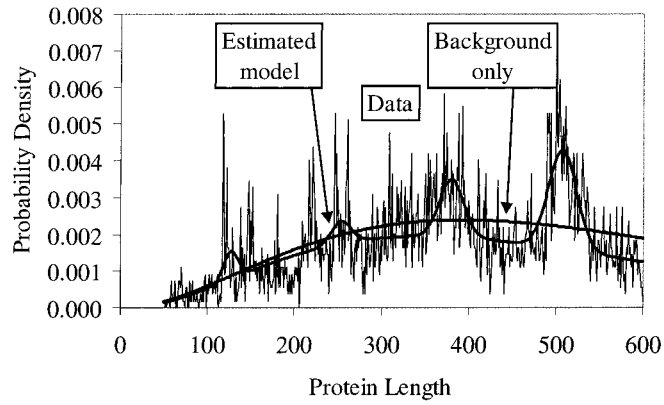

FIG. 7. Estimated probability density of eukaryotic enzyme lengths, nonredundant data set. The data are shown with the estimated model and the constrained model consisting of background only without any peaks.

(Fig. 6) has a peak at approximately 125 aa, as was seen in Figure 4 for the entire data set. Estimated statistical parameters and p values for both data sets (entire and nonredundant) sets are shown in Table 2, and the peaks are very highly significant (the p value is equivalent to a Z value of 20.3). The background and maximum likelihood models for the nonredundant data (Fig. 7) show strong similarity to the entire data set (Fig. 3). Altogether, these results suggest that both the entire and the nonredundant sets of eukaryotic enzymes have typical preferred protein (domain) unit length of $\sim$125 aa.

CONCLUSION

This study introduces an original spectral analysis approach, SAD, to find periodic signal components in raw experimental data distributions (histograms) of arbitrary shapes. The proposed methods allow detection and rigorous evaluation of possible periodic components contained in data distributions. The periods extracted are not constrained to be integer fractions of the interval over which the distribution is defined, thus providing a more continuous spectrum of the distribution. The SAD technique is, thus, a useful research tool with improved diagnostic value as compared with traditional analytic methods. The SAD approach was illustrated by analysis of the eukaryotic enzyme data sets. The estimated model, with a preferred protein (domain) unit size of approximately 125 aa among eukaryotic enzymes, is in accordance with earlier observations on eukaryotic protein sequence length (Berman et al., 1994; Svedberg, 1929) and protein structural domain (Wheelan et al., 2000) preferences.

ACKNOWLEDGMENTS

We are grateful to P. Edlefsen, G. Glusman, D. Haynor, L. Hood, A. Keller, P. Shannon, A. Smit, S. Stolyar, and T. Xie for valuable comments and L. Hood for support. A. Siegel holds the Grant I. Butterbaugh Professorship at the University of Washington. This work was supported by the Institute for Systems Biology Program for the innovative research and in part by the U.S. National Science Foundation grant no. PHY99-07949 (E.N.T.).

REFERENCES

BERMAN, A.L., KOLKER, E., and TRIFONOV, E.N. (1994). Underlying order in protein sequence organization. Proc Natl Acad Sci USA **70,** 4044–4047.

SVEDBERG, T. (1929). Mass and size of protein molecules. Nature (Lond) **123,** 871.

POHLMANN, K.C. (2000). *Principles of Digital Audio*, 4th ed. (New York: MacGraw-Hill).

KENDALL, M., and STUART, A. (1979). *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship*, 4th ed. (New York: MacMillan).

BAIROCH, A., and APWEILER, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res **28,** 45–48.

WHEELAN, S.J., MARCHLER-BAUER, A., and BRYANT, S.H. (2000). Domain size distributions can predict domain boundaries. Bioinformatics **16,** 613–618.

Address reprint requests to:
Dr. Eugene Kolker
The Institute for Systems Biology
1441 North 34th Street
Seattle, WA 98103-8904

E-mail: ekolker@systemsbiology.org