Distribution of Indel Lengths

Bin Qian[1] **and Richard A. Goldstein**[1,2*]
[1]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **Protein sequence alignment has become a widely used method in the study of newly sequenced proteins. Most sequence alignment methods use an affine gap penalty to assign scores to insertions and deletions. Although affine gap penalties represent the relative ease of extending a gap compared with initializing a gap, it is still an obvious oversimplification of the real processes that occur during sequence evolution. To improve the efficiency of sequence alignment methods and to obtain a better understanding of the process of sequence evolution, we wanted to find a more accurate model of insertions and deletions in homologous proteins. In this work, we extract the probability of a gap occurrence and the resulting gap length distribution in distantly related proteins (sequence identity < 25%) using alignments based on their common structures. We observe a distribution of gaps that can be fitted with a multiexponential with four distinct components. The results suggest new approaches to modeling insertions and deletions in sequence alignments. Proteins 2001;45:102–104.**
© 2001 Wiley-Liss, Inc.

Key words: **sequence alignment; insertion and deletion; gaps; protein evolution; dynamic programming**

INTRODUCTION

For the past decades, protein sequence alignment has become a widely accepted first step in the study of newly found protein sequences. Popular sequence alignment software include FASTA,[1] BLAST,[2,3] SSEARCH,[1,4] and more recent iterative methods such as PSI-BLAST.[5] Based on dynamic programing algorithms, these methods can identify homologies between the target protein and those in currently-available protein sequence databases. Identification of such homologs can provide valuable information regarding the structural and functional properties of the target sequences.

In most of these sequence comparison programs, a scores θ will be calculated on the basis the aligned pairs of amino acid, and affine gap penalty will be used to assign scores to insertions and deletions where residues from one sequence were aligned with “gaps” in another sequence:

$$\theta = \sum_{i,j} m(A_i, A_j)\gamma_{i,j} + \sum_{k} (\gamma_{\text{gap-I}} + \gamma_{\text{gap-E}}(n_k - 1)) \qquad (1)$$

where $m(A_i, A_j)$ refers to the number of times that amino acid type A_i is aligned with amino acid type A_j, n_k is the length of gap k, $\gamma_{i,j}$ represents the contribution to the score for any amino acid match or mismatch, and $\gamma_{\text{gap-I}}$ and $\gamma_{\text{gap-E}}$ represent the penalty for opening a gap and extending the gap, respectively. The use of this two-parameter affine gap penalty is motivated by two assumptions: first, residues in the gap are randomly distributed; second, the probability of the gap occurring at any location is a geometric distribution of the form

$$P(n_k) = P_g\left(\frac{\exp(-1/\lambda)}{1 - \exp(-1/\lambda)}\right)\exp(-n_k/\lambda) \qquad (2)$$

where P_g is the probability of opening a gap. Thus, the probability Q_k of the occurrence of gap k with length of n_k found opposite a stretch of amino acids $\{A_i\}$ in the corresponding homologous protein is

$$Q_k = P(n_k) * \prod_{i=1}^{n} q(A_i) \qquad (3)$$

where $q(A_i)$ is the probabilities of the random residues occurring in the homolog.

We generally are interested in the log of the ratio of the probability of the observed match to the probability that such a match would occur at random in nonhomologous proteins. When we divide Q_k with the residues expected to occur in a random match to a nonhomologous protein, the $q(A_i)$ cancel out. Taking the logarithm yields a contribution to the score of gap k equal to

$$\theta_k^{\text{gap}} = \log(P_g) + \log\left(\frac{\exp(-1/\lambda)}{1 - \exp(-1/\lambda)}\right) - (1/\lambda)n_k \qquad (4)$$

which corresponds to equation 1 if $\gamma_{\text{gap-I}} = \log[P_g/(1 - \exp(-1/\lambda))] - 2/\lambda$ and $\gamma_{\text{gap-E}} = -1/\lambda$.

The affine gap penalty captures the qualitative sense that, evolutionarily, it is harder to open a gap than to extend one. On the other hand, the affine gap assumption is an obvious simplification of the real world in which there are different mechanisms for a sequence to be inserted or

Grant sponsor: College of Literature, Science, and the Arts; Grant sponsor: Program in Protein Structure and Design; Grant sponsor: Horace H. Rackham School of Graduate Studies; Grant sponsor: National Institutes of Health; Grant number: LM0577; Grant sponsor: National Science Foundation; Grant number: BIR9512955.

*Correspondence to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@umich.edu

Received 3 November 2000; Accepted 7 May 2001

© 2001 WILEY-LISS, INC.

to be deleted. For example, the "bulge" of one strand may cause deletion in the daughter strand during the DNA replication; the repair of the "bulge" may also cause deletion of a DNA strand. Crossovers and transpositions can introduce even larger insertions and deletions.[6] Using a single expression to represent the multiplicity of complicated events that have accumulated over a long evolutionary time can have a significant impact on the accuracy of a scoring scheme. This finding suggests that we can improve sequence alignment methods by implementing a more accurate gap penalty scheme.

To determine an appropriate scoring scheme, we want to examine the real evolutionary pattern of insertions and deletions, which can also help us understand the underlying evolutionary processes. To do this, we take advantage of the availability of sets of structurally similar proteins that have been aligned on the basis of their shared structures, in our case the Family of Structurally Similar Proteins (FSSP) database.[7] Although the structurally based alignments do not necessarily reflect the correspondence between evolutionarily related amino acids in the two sequences, it is still the best available approach for distant homologies. We performed a maximum entropy analysis of the distribution of insertions and deletions in distantly related homologs, finding that the distribution fits a multiexponential expression with four distinct components. Possible applications to sequence alignment methods are also discussed.

MATERIALS AND METHODS

Theory

We use the structurally aligned set of proteins to generate a distribution of the length of observed insertions and deletions of the form $\{x_n\}$, where x_n is the number of observed gaps of length n. We are then interested in representing the observed distribution with a model, which we will represent with M. M includes all of the various parameters to be determined based on the data.

According to Bayes' theorem, the conditional probability of the model given the data D and underlying assumptions U, $P(M|D, U)$ is proportional to the product of $P(D|M, U) * P(M|U)$ where $P(D|M, U)$ is the probability of the observed data resulting given the model and the underlying prior assumptions, and $P(M|U)$ is the prior probability of the model given only the assumptions. The most likely model is the one that maximizes $P(D|M, U) * P(M|U)$, or equivalently the sum of the logarithms of these two terms.

We assume the data $\{x_n\}$ fits a sum of N exponentials of the form.

$$\text{Est}(x_n) = \sum_i^N A_i \exp(n/\lambda_i) \tag{5}$$

where $\text{Est}(x_n)$ is the expected value of x_n. For the a priori assumption that the observed gaps are randomly distributed among the various components in the multiexponential, the logarithm of $P(M|U)$ is proportional to the entropy S given by

$$S = -\sum_i p_i \log p_i \tag{6}$$

where p_i is the fraction of all gaps in component i, equal to $p_i = (A_i\lambda_i)/\Sigma_{i'}(A_{i'}\lambda_{i'})$.[8] Assuming random and sufficiently large statistics, the logarithm of $P(D|M, U)$ is proportional to the conventional $-\chi^2$. In this framework, we then want to minimize

$$\chi^2 - a * S \tag{7}$$

where a represents the relative weight of the two terms and our confidence in the presumed prior.

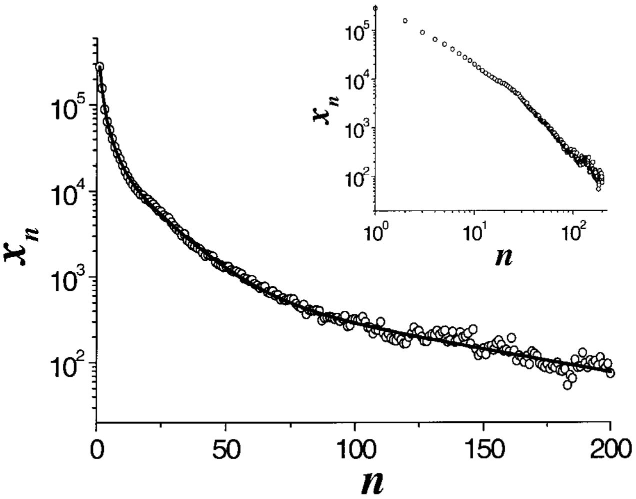

Fig. 1. Log plot and log-log plot of observed structure-based gap length distribution, compared with a quadruple-exponential fit (Eq. 8).

The probability of gap occurrence P_g is calculated by counting the total number of gaps and dividing by the total number of locations in the sets of aligned proteins where a gap could start, that is, the locations not either in gaps or immediately following a gap.

Database Preparation

There are several structure based categories of protein space,[9] but only FSSP[7] has explicit structure alignment profiles. In each FSSP set, a group of structurally related sequences are aligned with one representative sequence, with explicit alignments profiles. We use a set of 1959 FSSP protein sets and chose those sequence pairs with identity $\leq$ 25% in each aligned set, for a total of 167,712 locally aligned sequences. All gaps with length $<$ 200 residues were tabulated. The probability of gap occurrence for sequence identity $<$ 25% sequences is $P_g = 0.030$. Figure 1 shows the gap length distribution of FSSP structural alignments.

RESULTS

As shown in Figure 1, the gap length distribution cannot be represented as a straight line in a log plot, indicating that the distribution cannot be represented with a simple exponential. As described above, we used a maximum entropy formalism to fit the data to a multiexponential.

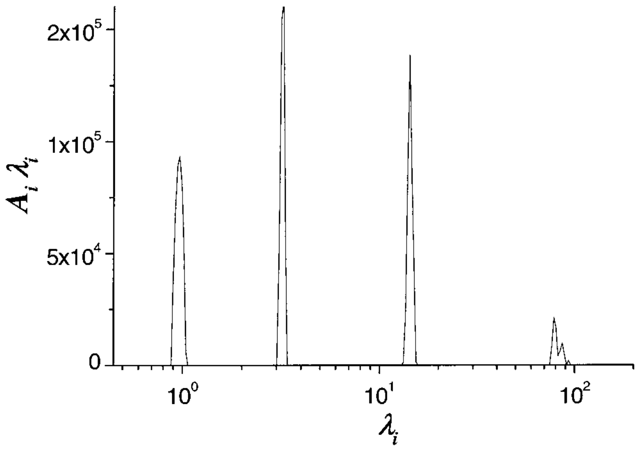

Fig. 2. The distribution of exponential terms obtained with a maximum entropy fit, for $a = 1$.

Our initial formula contained 600 exponential terms with values of λ_i geometrically distributed between 0.021 and 400. We minimized $\chi^2 - a * S$ as described in Eq. 7 with $a = 1$, 100, and 1,000. All three optimizations yield four distinct peaks in the distribution of $A_i\lambda_i$ values, as shown in Fig. 2 for $a = 1$. On the basis of these results, we fit the data to a simple quadruple exponential. The optimal fit was obtained with

$$\begin{aligned} P(n) = {} & 1.027 \cdot 10^{-2}\exp(-n/0.96) \\ & + 3.031 \cdot 10^{-3}\exp(-n/3.13) \\ & + 6.141 \cdot 10^{-4}\exp(-n/14.3) \\ & + 2.090 \cdot 10^{-5}\exp(-n/81.7) \end{aligned} \tag{8}$$

where $P(n)$ is the probability of gap with length n.

We can approximate this quadruple exponential in a piecewise manner as four separate exponentials appropriate for different size gaps. The corresponding gap penalties can be set by using Eq. 4 or through optimization.[10] In this way, we can implement these distinct insertion and deletion statistics into standard dynamic programming routines as a set of affine gaps with only a small increase in computational complexity, keeping the favorable scaling relationships of standard dynamic programing.

DISCUSSION

The gap length information generated from the structural alignment of these protein sequences verified our suspicion of the validity of an affine gap penalty. As shown above, the gap length distribution is not a single exponential as assumed by an affine gap penalty. Instead, it can be satisfactorily fit by a quadruple exponential. It is possible that more data would provide better resolution of the distribution of gap lengths. But the quadruple affine gap can be easily incorporated into standard dynamic programing methods with only a moderate increase in the computational complexity.

The distribution of gap lengths may represent different mechanisms of insertions and deletions, with particular length scales. For example, DNA mispairing will only affect a small portion of DNA, generating short indels. Conversely, insertions and deletions caused by crossover or transposition events generally include longer sequences. These different approaches probably have different rates, generating the observed complex pattern of gap lengths.

Our quadruple exponential model was generated from distantly related protein sequences, so it can be most appropriately used in aligning dissimilar sequences. If these four terms in fact represent different evolutionary processes, it is likely that the quadruple exponential model with appropriatly adjusted gap penalties can be used in sequence alignments for other evolutionary distances.

Benner et al.[11] inferred a power-law distribution of gap lengths. According to this model, the distribution of gap lengths should be linear on a log-log plot. As seen on the log-log plot in Figure 1, although this holds for shorter gaps (λ), this is not consistent with our data for longer gap lengths. It is not possible for our data to be represented by a sum of such power-law distributions because this would result in a concave-up curve in a log-log plot in contrast to the concave-down shape in Figure 1. The concave-down shape also argues against a logarithmic gap penalty even for shorter gaps, because the mechanism in the shorter gap region would be expected to extend to the longer gap length region by evolutionary accumulation, again resulting in a concave-up shape.

ACKNOWLEDGMENTS

We thank Maricel Kann for helpful discussions and Todd Raeker and Michael Kitson for computational assistance.

REFERENCES

1. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448.
2. Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. J Mol Biol 1990;215:403–410.
3. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol 1996;215:460–480.
4. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227:1435–1441.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DL. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
6. Li WD, Graur D. Fundamentals of molecular evolution. Sunderland: Sinauer; 1991.
7. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
8. Gelman A. Bayesian data analysis. London: Chapman and Hall; 1995.
9. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7:1099–1112.
10. Kann M, Qian B, Goldstein RA. Optimization of a new score function for the detection of remote homologs. Proteins 2000;41: 498–503.
11. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein database. Science 1992;256:1443–1445.