



Estimating the diversity of peptide populations from limited sequence data

Lee Makowski^{1,*} and Alexei Soares²

¹Bioscience Division, Argonne National Laboratory, 9700 S. Cass Avenue Argonne, Illinois 60439, USA and ²Biology Department, Brookhaven National Laboratory, Upton, New York 11973, USA

Received on July 5, 2002; revised on September 20, 2002; accepted on September 28, 2002

ABSTRACT

Motivation: Combinatorial libraries of peptides such as those displayed on the surface of a bacteriophage particle have become widely used tools for characterizing protein–protein and protein–small molecule interactions. The quality of a library frequently depends on its completeness, or diversity—the proportion of possible sequences actually present in the library. The diversity of these libraries is frequently quoted on the basis of phage titers that provide little information about their completeness.

Results: Here, an analytical expression for diversity is introduced and a method for estimating the diversity of a peptide library from the sequences of a limited number of the members of the library is demonstrated. The diversities of a number of computationally constructed and actual peptide libraries are estimated using this method.

Contact: lmakowski@anl.gov

INTRODUCTION

Phage-displayed peptide libraries are routinely used for mapping epitopes, identifying peptide ligands, mapping protein–protein interactions and identifying the specificities of enzymes (e.g. Rodi *et al.*, 2001). The utility of combinatorial libraries is directly dependent on the amino acid sequence diversity of the library. Numerous investigators have attempted to analyze the diversity of peptide libraries but there is, as of yet, no rigorous quantitative way to measure sequence diversity.

As a surrogate for a true measure of diversity, the number of independent clones in the library and the number of copies of each peptide are sometimes quoted as a measure of library complexity (e.g. Noren and Noren, 2001). Scott and Smith (1990) calculated the probability of peptides being present in a library of 2.3×10^6 clones assuming Poisson statistics and equal probability of occurrence for all possible clones. Cwirla *et al.* (1990) recognized that the apparent diversity of a peptide library

will be limited by the fact that each of the 20 amino acids is coded for by different numbers of codons. They further observed that most amino acids occurred at most positions in their hexapeptide library, leading them to conclude that viral morphogenesis did not impose severe constraints on the diversity of their library. The effect of viral morphogenesis is, in fact, observable and significant but, consistent with their observations, not severe (Rodi *et al.*, 2002). Cwirla *et al.* (1990) further recognized that populations of selected peptides exhibited more diversity at some positions than at others, but did not attempt to quantitate this observation. DeGraff *et al.* (1993) carried out a more extensive analysis of diversity in a phage displayed library of random decapeptides. They analyzed the sequence of 52 clones selected at random from a population of 2×10^6 individual clones and demonstrated that the frequency of amino acid occurrence in this library had a rough correlation with that expected from the number of codons corresponding to each amino acid. They further showed that 250 of the 400 possible dipeptides were present in the 52 decapeptides selected. Since only 468 dipeptides are included in this limited population, this observation is not significantly different from that expected from random sampling.

Each of these analyses was motivated by the need to measure the diversity (or complexity) of a peptide library. Although an accurate measure of diversity could be made by sequencing millions of peptides and recording the number of times each peptide occurs in the library, this procedure is experimentally tedious and expensive. Furthermore, even if it were possible to obtain this amount of sequence data, no general quantitative measure of diversity is available to evaluate the relative diversity values of different libraries.

There are two possible approaches to the definition of diversity—‘technical diversity’, or completeness, corresponding to the percentage of possible members of a population that exist at any copy number within a population; and ‘functional diversity’, which takes into account the copy numbers of each distinct member of the

*To whom correspondence should be addressed.

population (Rodi *et al.*, 2002). In the latter case, if the copy numbers of the members present in the population are dramatically different, the diversity is intrinsically lower. Experiments that utilize limited sequence information to estimate peptide population diversity cannot provide accurate estimates of completeness since very rare members of the population will inevitably go unsampled. However, as outlined below, limited sequence information is capable of estimating the functional diversity of a peptide library.

Here, we propose an analytical expression for the diversity of a population of peptides, demonstrate that this expression is consistent with intuitive expectations for the properties of population diversity, and provide a means to calculate the diversity on the basis of a limited number of peptide sequences. Inherent in this definition for diversity is the observation that diversity is not simply a measure of how many peptide sequences exist within a population, but is also dependent upon the relative abundance of each peptide within that population. Real phage displayed peptide populations invariably contain unequal numbers of different peptides and any useful measure of sequence diversity must take this into account.

SYSTEMS AND METHODS

A quantitative expression for sequence diversity

Consider the process by which a single member of a population is selected from that population. If the population has N members, each of which is present in equal abundance, then the probability of any one member being chosen is $(1/N)$. If that population cannot, theoretically, be greater than N , then we say that the population has a diversity, d , equal to 1.0. In other words, all of the theoretically possible members of this population are equally abundant in the population.

Consider a second population containing equally probable members of half ($N/2$) of the possible sequences, but none from the other half. Intuitively, the diversity of this population is 0.5. In other words, 50% of the theoretically possible members are present in equal numbers in this population. Consider a third population containing only a single member of the N possible members. The diversity of this population is $(1/N)$. These examples constitute ideal cases in which all members that are present are present in equal abundance. These are useful initiation points because their diversity is intuitively obvious and any quantitative expression for diversity must be consistent with these intuitive results. However, an expression for diversity must also be applicable to peptide populations in which different members are not equally abundant.

Figure 1 contrasts three populations, each containing at least a few copies of all N possible members. In population 1, all members are present in equal number. In pop-

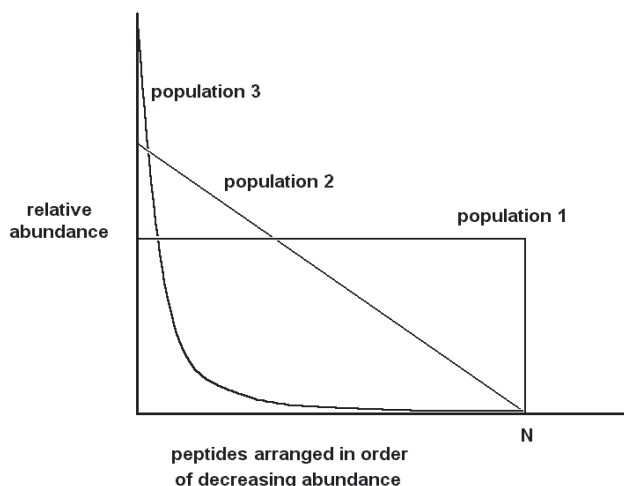


Fig. 1. Relative abundance of peptides in three hypothetical populations of peptides. In population 1 all peptides have equal abundance. In population 2, there is a linear decrease in abundance throughout the population. In population 3 a relatively small number of peptides are present at relatively high abundance levels.

ulation 3, a small number of members are present in relatively high abundance. When members of these populations are drawn at random, the probability of drawing multiple copies of a specific peptide will be greatest in population 3 and lowest in population 1. Consequently, population 3 will appear to be less diverse than population 1. This is true even though both populations have at least a few members of every possible peptide. These examples suggest three criteria for an acceptable quantitative measure of sequence diversity: (i) the measure of diversity should be consistent with the intuitive estimates of diversity made in the examples outlined above; (ii) for any two populations with the same 'diversity', the probability of choosing the same member twice in random selection from the two populations should be the same; and (iii) this probability should be greater in a population with lower diversity than in a population with higher diversity.

An expression for diversity that fits these criteria is as follows: For a population in which there is a theoretical maximum of N possible members, the diversity, d , is defined as

$$d = 1/(N \sum_k p_k^2) \quad (1)$$

where the sum is over all possible members, k , and p_k is the probability of the k th member being selected in any random selection from the population—a direct measure of the relative abundance of the peptides. Note that for any population, $\sum_k p_k = 1$.

Several examples demonstrate that this definition satisfies the intuitive criteria for simple populations in which all peptides that are present are present in

equal numbers. If all possible members are present in equal numbers (equal probability of being chosen), then $p_k = 1/N$ for all k . The sum in equation (1) is then equal to $N(1/N^2) = (1/N)$, and $d = 1.0$ as expected. If half ($N/2$) of the members are present in equal numbers and the other half are missing, then for half the population, $p_k = 2/N$, and for the other half, $p_k = 0$. The sum in equation (1) is then $(N/2)(2/N)^2 = 2/N$, and it follows that $d = 0.5$ (as intuitively expected).

What about the case where all members are present but half of the members are present at twice the abundance of the other half? The criterion that $\sum_k p_k = 1$ can be used to derive the probabilities of the two subpopulations, leading to the conclusion that the members of the more probable half have a probability of selection of $(4/3N)$; whereas the probability of selection of any one of the less probable members is $(2/3N)$. Substituting these values into equation (1), the diversity of this population is calculated to be 0.9. This is consistent with the intuition that this population is less diverse than a population in which all members are equally abundant but more diverse than a population in which only half the members are present. How is it related to a simple population in which 90% of the members are present in equal numbers and the other 10% are absent? This population also has a diversity of 0.9. As will be derived below, in a random selection process, the probability of selecting the same member twice is the same for these two populations, and, in fact, for all populations having a diversity of 0.9.

The measure of diversity, as defined in equation (1) is closely related to the effective number of alleles introduced by (Kimura and Crow, 1964) and used in population genetics (Nei, 1987). In that body of the literature, the effective number of alleles differs from diversity as defined here only in that it is scaled by library (population) size.

The probability of selecting the same peptide twice from a population

Consider an experiment where peptides are selected one at a time at random from a very large population with diversity = d . Assume that at some point, m unique peptides have been chosen. If one more peptide (k) is selected, the probability, $P(\text{match})$, that this peptide is exactly identical to one of the m unique peptides already chosen is

$$P(\text{match}) = \sum_k P(\text{peptide } k \text{ being chosen}) \times P(\text{peptide } k \text{ already being in the pool})$$

where the sum is over all possible peptides, k . The first term in the product is p_k the second is (for the case of $mp_k \ll 1$) is mp_k . Consequently,

$$P(\text{match}) = \sum_k p_k mp_k = m \sum_k p_k^2$$

Since, from equation (1), $\sum_k p_k^2 = 1/Nd$, it follows that $P(\text{match}) = m/Nd$.

Consequently, $P(\text{match})$ is the same for all populations with the same diversity, d .

ALGORITHM

Calculating sequence diversity

The expression for diversity in equation (1) is of limited practical utility because it requires a sum over every possible peptide in a population. For a 12mer peptide population, that sum requires the estimation of 4×10^{15} probabilities, an impractical operation. This problem can, however, be readily circumvented for populations in which the probability of an amino acid occurring at any position is independent of the identity of amino acids at other positions:

For a population of peptides, define p_{ij} as the probability of amino acid j occurring at position i in the peptide. If the occurrence of an amino acid at one position in a peptide is independent of the identity of another amino acid at another position, then the probability of peptide k occurring is simply the product $p_k = \prod_i p_{ij}$, where the product is over all positions in the peptide and the p_{ij} reflect the probability of occurrence of amino acid j at position i . In this case, the sum in equation (1) reduces to

$$\sum_k p_k^2 = \prod_i (\sum_j p_{ij}^2)$$

where the sum on the right hand side of this equation is over all possible amino acids, j , and the product is over all positions, i . Substituting into equation (1)

$$d = 1/(N \prod_i (\sum_j p_{ij}^2)). \tag{2}$$

Generally, the probabilities, p_{ij} , are not known, but can be estimated from the frequencies, f_{ij} , of occurrence of each amino acid at each position. For peptides M amino acids in length, $N = 20^M$, and:

$$d = 1/(N \prod_i (\sum_j f_{ij}^2)) = 1/(20^M \prod_i (\sum_j f_{ij}^2)). \tag{3}$$

This result is readily demonstrated for simple cases. For instance, for a complete population of equally probable peptides M amino acids in length, all terms p_{ij} are equal to $(1/20)$; and, since there are 20 amino acids, j , the sum $(\sum_j f_{ij}^2)$ is equal to $1/20$ for all positions, i . It follows that the product is equal to $(1/20^M)$. Since there are 20^M possible peptides in a population M amino acids in length, the diversity is calculated to be 1.0 as it must be.

Estimating uncertainty in the calculation of diversity

A second challenge in the calculation of the diversity of a population is determining the accuracy of the diversity

estimate based on observed frequencies. As will be demonstrated below, the direct use of equation (3) leads to significant underestimate of diversity when less than about 500 peptide sequences are known. This systematic error can, however, be corrected by taking into account the effect of uncertainties in the estimation of peptide frequencies at each position.

The estimate of frequency of occurrence of amino acid, j , at any position, i , has, assuming a Poisson distribution, a standard deviation of (e.g. Bulmer, 1979),

$$\sigma(f_{ij}) = (f_{ij}(1 - f_{ij})/n_{\text{pep}})^{1/2}, \tag{4}$$

where n_{pep} is the number of peptides used to estimate the frequencies, f_{ij} .

The systematic error in the estimation of diversity given in equation (3) is due to the fact that diversity involves a sum over the *squares* of the frequencies. Any error in the frequency (over-estimate or under-estimate) leads to an overestimate of the sum, and a resultant underestimate of the diversity.

When errors, $\varepsilon(f_{ij})$, are present in the estimates of frequencies, f_{ij} , the estimated diversity is

$$d = 1/(20^M \prod_i [\sum_j (f_{ij} + \varepsilon(f_{ij}))^2]). \tag{5}$$

When the square in the denominator is taken, the cross terms involve both positive and negative errors and will sum to approximately zero because of the restriction that $\sum_j f_{ij} = 1$ assuming that the sign of the error is not correlated with the magnitude of the frequency. Results presented in Figures 2 and 3 suggest that this assumption is justified. Consequently, equation (5) reduces to

$$d = 1/(20^M \prod_i [\sum_j (f_{ij}^2 + \varepsilon^2(f_{ij}))]). \tag{6}$$

In this equation, the error terms, $\varepsilon^2(f_{ij})$, are all positive giving rise to a significant underestimate in diversity when the calculation is based on small numbers of sequences (less than about 500). Compensation for this effect can be accomplished by *subtracting* the expected error from the sum in the denominator. Then, the estimate of the diversity, d_e , becomes:

$$d_e = 1/(20^M \prod_i [\sum_j (f_{ij}^2 - \sigma^2(f_{ij}))]) \tag{7}$$

where the error term has been replaced by the divergence. As will be demonstrated below, this expression provides reasonable estimates for the diversity of a library when the sequences of as few as 50 peptides are available.

The propagation of estimated errors in frequencies, f_{ij} , through equation (7) can be calculated using standard formulae to obtain a standard deviation for the estimate of diversity equal to

$$\sigma(d_e) = \frac{\left\{ 2 \sum_i \left[\frac{(\sum_j f_{ij}^2 \sigma^2(f_{ij}))^{1/2}}{\sum_j f_{ij}^2 - \sum_j \sigma^2(f_{ij})} \right]^2 \right\}^{1/2}}{20^M \prod_i (\sum_j f_{ij}^2 - \sum_j \sigma^2(f_{ij}))}. \tag{8}$$

Estimated Diversity in a 61 Codon Library

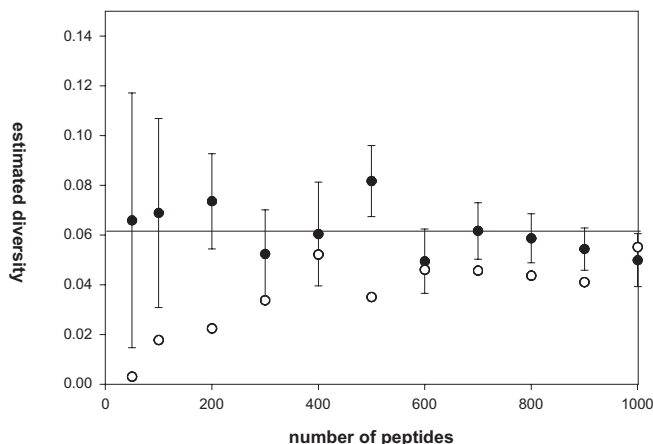


Fig. 2. Comparison of the diversity of a population of 12mer peptides as calculated from equation (3) using a 61 codon genetic code (horizontal line corresponding to 0.0606) with estimates of diversity calculated from different sized populations of peptides generated by random selection with a 61 codon code. The filled circles correspond to the diversity estimated using equation (7) after taking into account systematic errors due to sampling. Error bars were calculated from equation (8). The open circles correspond to the same estimates without taking the systematic errors into account.

Estimated Diversity in a 32 Codon Library

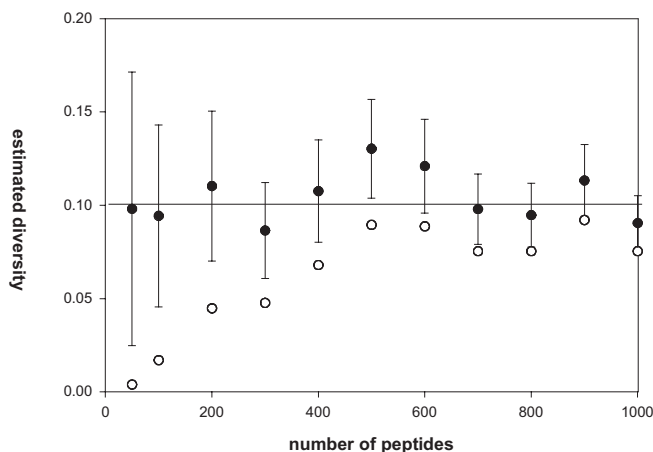


Fig. 3. Comparison of the diversity of a population of 12mer peptides as calculated from equation (3) using a 32 codon genetic code (horizontal line corresponding to 0.1006) with estimates of diversity calculated from different sized populations of peptides generated by random selection with a 32 codon code. The filled circles correspond to the diversity estimated using equation (7) after taking into account systematic errors due to sampling. Error bars were calculated from equation (8). The open circles correspond to the same estimates without taking the systematic errors into account.

This provides a measure of the random error in the estimation of diversity, d_e , as calculated in equation (7).

IMPLEMENTATION

Diversity of random codon libraries

Construction of peptides through the random selection of codons from a library of 61 possible codons does not result in a population with equal frequencies of all possible amino acids. Three amino acids have six codons; five have four codons; one has three; nine have two and two have one codon. These numbers can be used to calculate frequencies in an infinitely large population, the diversity of which can be calculated using equation (3). It follows, that for peptides M amino acids in length, a population of peptides constructed through the random selection of a set of 61 codons has a diversity of $(0.7919)^M$. For a library of 12mers, for instance, this corresponds to a diversity of 0.0606. In other words, in random selection from this library, the probability of selecting the same peptide twice would be the same as selecting from a 12mer library in which only 6.06% of all possible peptides are present in equal proportions.

Populations of 12mers were generated computationally using random selection from a 61 codon genetic code, and diversity of these populations was estimated using both equations (3) and (7). The results of these calculations are presented in Figure 2. The estimates obtained using equation (3) are represented by open circles and severely underestimate the diversity of the population when small numbers of peptides are used in the calculation. Even for large subpopulations, these estimates only very slowly approach the expected value of 0.0606. The estimates obtained using equation (7), corrected for the effect of random fluctuations in the observed frequencies are given by filled circles in this figure, and provide a reasonable estimate (± 0.01) for diversity even when based on as few as 50 peptides. For numbers of peptides below 50, the estimates given by equation (7) diverge rapidly.

Similar calculations can be carried out using a 32 codon genetic code (commonly used in the construction of phage displayed peptide libraries in which a suppressor host allows one termination codon to be translated as a glutamate). Using this coding, three amino acids have three codons; six have two codons and 11 have one codon. From equation (3), the diversity of a library constructed in this way is $(0.8258)^M$. For a 12mer library, the resulting diversity is 0.1006. Populations of 12mers were generated computationally using random selection from a 32 codon genetic code and diversity was calculated as in the case of a 61 codon library. The results of these calculations are in Figure 3, and are very similar to the results for a 61 codon library, except for the increase in diversity to match that expected for a 32 codon construction.

Diversity of proteome sequences

The diversity of peptides selected from genomic sequences can also be calculated using equation (3) directly from the relative abundances of amino acids in a genome. For instance, using the frequencies of occurrences of amino acids in human and *E. coli* genomes, the diversities calculated from equation (3) are $(0.850)^M$ and $(0.818)^M$, respectively. Note that both genomes exhibit a sequence diversity in excess of that expected for random selection from a 61 codon genetic code. That is due to the fact that the usage of amino acids by either organism is more uniform than would be expected for random usage of 61 codons. From the point of view of diversity of sequence, the more uniform usage of amino acids, the higher the diversity of sequence.

Diversity of phage displayed peptide libraries

The diversity of peptide libraries displayed on the surface of phage particles can be estimated from limited sequence data using equation (7) as was carried out for the computationally generated libraries in part (A) above. Table 1 provides a few examples of these calculations. The diversity per amino acid quoted in the Table provides a measure of the degree to which the biology of the phage–host system constrains the sequences of the library. The p8 libraries are the most constrained, as might be expected given the rigorous structural and metabolic constraints on the nature of the inserts allowed near the amino terminus of p8 (Kishchenko *et al.*, 1994; Rodi and Makowski, 1997). The NEB 7mer library consists of seven random amino acids coded by 32 codons and inserted between a pair of cysteines that normally form an intramolecular disulfide bond displayed at the amino terminus of the mature p3 of M13. This library exhibits less censorship from the biology of the phage–host system than the p8 library as measured by the diversity per amino acid, but is more censored than the 12mer library which is an unconstrained library of inserts at the amino terminus of M13 p3 (also constructed with 32 codons). The relative diversities of these libraries is as would be expected given the details of the virus–host system in which they were constructed. The diversity per amino acid in the 12mer library approaches that expected for a random library constructed from a 61 codon code, but is still significantly less diverse than expected for random selection from a 32 codon code.

DISCUSSION AND CONCLUSION

The method for estimating diversity of peptide populations outlined here provides a new tool for evaluating the quality of peptide libraries. The estimation of diversity for computationally constructed populations of peptides demonstrates the accuracy of the diversity estimates as

Table 1.

Source	Naa	Diversity	Diversity per aa*	# Peptides
Petrenko	6	0.0750	0.65	108
NEB	7	0.0797	0.70	99
NEB	12	0.0435	0.77	101

Naa is the number of randomized amino acid positions used in the calculation. Diversity per amino acid refers to the average diversity per amino acid position where $(\text{diversity per amino acid})^{\text{naa}}$ is the total library diversity quoted. # peptides refers to the number of peptides used in the calculation.

Sources: The Petrenko library is described in Petrenko *et al.* (1996) and amino acid sequences of the members were provided by Dr V.Petrenko. It is an M13 p8 library in which 8 amino acids were randomized. The first and last codons were only partially randomized, and were ignored for this calculation. The remaining 6 amino acids were constructed using a 32 codon genetic code. The NEB libraries refer to libraries commercially available from New England Biolabs (Ph.D.-c7 and Ph.D.12). Sequences were obtained as described in Rodi *et al.* (2002).

defined here. In the case of both libraries where the diversity of a very large population could be calculated exactly, the estimates from peptide populations generated randomly from these populations reflect accurately (within $\pm 1\%$) the expected diversity. This validates the exact calculation of diversity as given in equation (3) and the estimate of diversity given in equation (6).

The results reflect a diversity which, on first reflection, may seem low. Is it really possible that random selection of codons would result in populations that behave as if their diversity were only 6–10%? Consider the fact that a complete library of 12mers contains 4.096×10^{15} members. A library with a diversity of 0.0606 will behave as though it contains 2.48×10^{14} members in equal proportions; a library with a diversity of 0.1006 will behave as though it contains 4.21×10^{14} members. These are still highly complex populations and any selection from them will have a huge diversity of biochemical and binding properties.

The diversity calculated here is a reflection of the diversity of the population from which the peptides have been selected. It does not speak to the size of the population from which the peptides have been selected. For instance, a small library of clones corresponding to only 10^4 members could exhibit a diversity representative of a population much larger than 10^4 . The calculation cannot estimate the number of independent clones, rather it reflects the statistical properties of the population from which the selected peptides have been obtained.

It is interesting to compare the diversities of computationally constructed libraries with those calculated from the amino acid abundance in the human and *E. coli* genomes. As calculated using equation (3), the diversity of peptide segments of length M selected at random from

the human genome will be approximately $(0.85)^M$, which equals 0.1421 for 12mers ($M = 12$). For comparison, the diversity estimated from equation (7) using 500 12mer peptides selected at random from the human genome was 0.13. Similar calculations were carried out for the *E. coli* genome, resulting in a calculation of diversity equal to $(0.8183)^M$, or 0.0901 for 12mers. The estimate from a single population of 500 12mers selected at random from the *E. coli* genome was 0.1060.

The diversity of the peptides used in the human genome is actually greater than those generated at random using either a 32 or 61 codon scheme. This is because the frequency distribution of amino acids in the human genome is closer to being uniform than it would be if it were generated using random codon selection from either 32 or 61 codons were used. The frequencies of amino acid occurrences in the human genome actually results in a more diverse sequence set. Similarly, the frequencies in the *E. coli* genome result in a diversity significantly greater than what would be generated by random selection from a 61 codon library, and nearly as great as what would be generated from a 32 codon library.

Diversity calculations using the analytical expressions introduced here can also be made for RNA or DNA sequences. Since only four bases make up a nucleotide sequence, reasonable estimates for sequence diversity can be made using as few as ten sequences.

In addition to estimating the diversity of populations of peptides or nucleotides as presented here, the diversity measure introduced here can be used to calculate the diversity as a function of position within a population of aligned sequences (Makowski, in preparation). This approach has significant potential as a quantitative measure of the relative degree of conservation of positions in families of proteins, and in coding or non-coding regions of genomes.

ACKNOWLEDGEMENTS

The authors thank F.J.Stevens for assistance with the error analysis and comments on the manuscript and Diane J. Rodi for helpful discussions. We also would like to thank an anonymous referee for pointing out the relationship between diversity, as defined here, and 'effective number of alleles' as used in population genetics. This work was funded by a grant from the Lucille P. Markey Foundation and a grant from the Office of Biological and Environmental Research, Department of Energy under Contract No. W-31-109-Eng-38.

REFERENCES

Bulmer, M.G. (1979) *Principles of Statistics*. Dover Publications, New York.

- Cwirla, S.E., Peters, E.A., Barrett, R.W. and Dower, W.J. (1990) Peptides on phage: a vast library of peptides for identifying ligands. *Proc. Natl Acad. Sci. USA*, **87**, 6378–6382.
- DeGraff, M.E., Miceli, R.M., Mott, J.E. and Fischer, H.D. (1993) Biochemical diversity in a phage display library of random decapeptides. *Gene*, **128**, 13–17.
- Kimura, M. and Crow, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.
- Kishchenko, G.H., Batliwala, L. and Makowski, L. (1994) Structure of a foreign peptide displayed on the surface of bacteriophage M13. *J. Mol. Biol.*, **241**, 208–213.
- Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Noren, K.A. and Noren, C.J. (2001) Construction of high-complexity combinatorial phage display peptide libraries. *Methods*, **23**, 169–178.
- Petrenko, V.A., Smith, G.P., X., Gong and T., Quinn (1996) A library of organic landscapes on filamentous phage. *Protein Engineering*, **9**, 797–811.
- Rodi, D.J. and Makowski, L. (1997) Transfer RNA isoacceptor availability contributes to sequence censorship in a library of phage displayed peptides. In *Proceedings of the 22nd Taniguchi International Symposium Nov. 18–21, 1996*.
- Rodi, D.J., Makowski, L. and Kay, B.K. (2001) One from column A and two from column B: the benefits of phage display in molecular-recognition studies. *Current Opinion in Chemical Biology*, **6**, 92–96.
- Rodi, D.J., Soares, A. and Makowski, L. (2002) Viral morphogenesis is the dominant source of sequence censorship in M13 combinatorial peptide phage display. *J. Mol. Biol.*, (in press)
- Scott, J.K. and Smith, G.P. (1990) Searching for peptide ligands with an epitope library. *Science*, **249**, 386–390.