Calculating the SNP-effective sample size from an alignment

Bernhard Haubold\textsuperscript{1,2} and Thomas Wiehe\textsuperscript{1,*}

\textsuperscript{1}Max-Planck-Institut f"ur Chemische Ökologie, Carl-Zeiss-Promenade 10, D-07745 Jena, Germany

Received on May 11, 2001; revised on June 27, 2001; accepted on August 7, 2001

ABSTRACT

Motivation: The number of Single Nucleotide Polymorphisms (SNPs) detectable in an alignment is a function of the length and the number of the aligned sequences. The latter is called sample size. However, a typical alignment, for instance obtained as a BLAST-search result of a query sequence against an EST database, does not evenly cover the query sequence. Therefore, it is usually not clear what the actual sample size is.

Results: We present a method to calculate the effective sample size, called $n_{\text{eff}}$, for a given BLAST alignment. This method takes into account that multiple coverage contributes only logarithmically to the SNP yield of a given sequence stretch. We show that the effective sample size $n_{\text{eff}}$ is usually much smaller than would be expected for a given amount of coverage and illustrate this with two typical examples.

Availability: The algorithm is implemented in NEFF, a program written in FORTRAN90 that is accessible at http://soft.ice.mpg.de/neff. From this site also the source, except for two subroutines protected by copyright, and a LINUX compiled executable can be downloaded.

Contact: twiehe@ice.mpg.de

INTRODUCTION

There is currently a lot of interest in the detection of Single Nucleotide Polymorphisms (SNPs; The International SNP Map Working Group, 2001; Venter et al., 2001). Their main application is in mapping genes, especially in species such as humans, where crossing experiments are not possible. SNPs are detected by comparing homologous DNA sequences and counting the variable positions. This number is commonly denoted by $S$. In the absence of selection there is a simple relationship (Watterson, 1975) between $S$ and $\theta$:

\[
\theta = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} \approx \frac{S}{\frac{1-\gamma}{\gamma n^2} + \log(n-1)},
\]

(1)

where $n$ is the number of sequences compared, and $\gamma$ is Euler’s constant.

The sample size, $n$, used in (1) depends on the assumption that the segment scanned for SNPs has been completely resequenced. However, SNPs are often inferred from fragments that are distributed unevenly across the gene (Figure 1). For example, the mouse green opsin gene yields sample sizes between 1 and 28 for different segments of its mRNA. Generally, to calculate $n_{\text{eff}}$ for a given mRNA of length $\Lambda$ we first subdivide the sequence into $k$ regions of length $l_j$, $j = 1, \ldots, k$. Within each such region the sample size is a well defined

---

\textsuperscript{*}To whom correspondence should be addressed.

\textsuperscript{2}Present address: LION Bioscience AG, Heidelberg, Germany.
Calculating the SNP-effective sample size

Fig. 1. BLASTN-result (Altschul et al., 1997) for the mouse green opsin mRNA (AH009854) aligned with the public mouse EST data base (dbEST release 040601).

integer, say \( n_j \). Then, for each region \( j \) with \( n_j > 1 \) the stacked sequence fragments are layed out linearly. What matters for our calculation is only the length and number of fragments; any sequence homology with the query sequence is disregarded. In this way a virtual sequence of length \( \Lambda + \lambda \) is constructed for which the sample size is exactly 2. It is related to the original multiple alignment by the fact that both yield the same average number of SNPs. For the linearization one has to observe that the sequence length \( L(n_j, l_j) \) contributed by a given region \( j \) to the number \( \Lambda + \lambda \) depends on both \( n_j \) and \( l_j \). From (1), and if \( n_j > 1 \), we obtain

\[
\theta(l_j + L(n_j, l_j)) = \theta l_j \left( \gamma + \frac{1 - \gamma}{n_j - 1} + \log(n_j - 1) \right),
\]

where again \( \theta = 4N\mu \), but \( \mu \) is now interpreted as the per nucleotide mutation rate. The left side of this equation gives the expected number of SNPs for sample size 2, the right side gives the expected number of SNPs for region \( j \) in the original multiple alignment. Solving for \( L(n_j, l_j) \) yields

\[
L(n_j, l_j) = l_j \left( \log(n_j - 1) - \frac{(1 - \gamma)(n_j - 2)}{n_j - 1} \right).
\]

Linearizing all regions \( j \) with \( n_j > 1 \) the virtual sequence has length \( \Lambda + \lambda \), where

\[
\lambda = \sum_{j=1; n_j > 1}^{k} L(n_j, l_j) - l_{(1)},
\]

and where \( l_{(1)} = \sum_{j=1; n_j = 1}^{k} l_j \) is the length of all regions for which \( n_j = 1 \). The sum on the right-hand side contains the ‘excess’ sequence obtained by the linearization process. From this number the combined length of those regions which do not contribute to the SNP yield, namely \( l_{(1)} \), needs to be subtracted.

Finally, and analogously to (2), we calculate the effective sample size \( n_{eff} \) in implicit representation:

\[
\theta(\Lambda + \lambda) = \theta \Lambda' \left( \gamma + \frac{1 - \gamma}{n_{eff} - 1} + \log(n_{eff} - 1) \right),
\]

where \( \Lambda' = \min(\Lambda, \Lambda + \lambda) \). An explicit number \( n_{eff} \) is obtained from (3) by a standard numerical routine (bisection method as implemented in rltbis; Press et al., 1996). Note that the minimum effective sample size according to (3) is \( n_{eff} = 2 \). However, if \( \lambda < 0 \), this minimum sample size refers only to a subsequence of the original sequence and is shorter than \( \Lambda \).

IMPLEMENTATION

NEFF takes as input the alignment output of a BLAST search in HTML or plain text format as produced, for instance, by the NCBI BLAST webserver (http://www.ncbi.nlm.nih.gov/BLAST) and returns the effective sample size, \( n_{eff} \), according to (3). The program is written in standard FORTRAN90 and can be accessed through a web interface at http://soft.ice.mpg.de/neff. On this site we also provide a stand-alone version of the software for Pentium II (or higher) processors and the Linux operating system. To allow for high throughput analysis, the input-file-name for the stand-alone program may contain wild cards so that an arbitrary number of files can be processed with a single call. In addition to the executable the website contains documentation, an example data file, and the source code, except for two subroutines taken from Press et al. (1996), which are protected by copyright.
APPLICATIONS

Human genome sequencing

In the Human Genome Issue of Nature a SNP was defined as a mismatch between two DNA sequences (Stoneking, 2001). Working from this definition, it might come as a surprise that the density of SNPs reported in the same issue of Nature was 1.8 times higher in genes than in intergenic DNA (The International SNP Map Working Group, 2001). However, as the authors of the SNP-map point out, the number of samples included in their search for polymorphisms was greater in genes than in intergenic regions (The International SNP Map Working Group, 2001). For intergenic stretches two sequences were compared, while for genes the sample size is unknown (D. Altschuler, personal communication).

When comparing two sequences, the number of SNPs per nucleotide is equal to the nucleotide diversity, \( \pi \). Given the overall genetic diversity of \( \pi = 7.51 \times 10^{-4} \), there is on average 1 SNP per 1/(7.51 \times 10^{-4}) = 1.33 kb of human genomic DNA (The International SNP Map Working Group, 2001). Notice that this differs from the reported density of SNPs on the human SNP-map (1/1.91 kb), because the coverage of the genome was not uniform and SNPs located in repetitive regions were discarded (The International SNP Map Working Group, 2001).

We estimated the relative genetic diversity in coding versus intergenic regions from the SNP density in exons (529 SNPs/Mb) and intergenic DNA (707 SNPs/Mb; Venter et al., 2001). This allowed us to calculate the effective sample size, \( n_{\text{eff}} \), used to determine the number of SNPs in coding regions. If the sample size for coding regions had been the same as for intergenic regions (i.e. 2), there would be 1 SNP per 1.33 \times 707/(529) = 1.78 kb. Knowing that the SNP-consortium found 1 SNP per 1.08 kb of coding sequence (The International SNP Map Working Group, 2001), we can write

\[
1.08 \approx \frac{\sum_{i=1}^{\gamma} \frac{1}{7}}{\sum_{i=1}^{\gamma} \frac{1}{7}} = 1 + \frac{1 - \gamma}{n_{\text{eff}} - 1} + \log(n_{\text{eff}} - 1)
\]

and find

\[ n_{\text{eff}} = 3.46. \]

BLAST search of EST database

When the mouse green opsin gene is compared to the public mouse EST database using BLAST, a ragged alignment is returned (Figure 1). Using (3) as implemented in NEFF, we find that each base pair is effectively covered \( n_{\text{eff}} = 2.92 \) many times. In contrast, the average sample size, obtained as the total length of query sequence and all ESTs and then divided by the length of the query sequence, in this example is \( n = 5.15 \).

DISCUSSION AND CONCLUSION

In this note we address the problem of comparing the potential for SNP detection between multiple local alignments of arbitrary raggedness. We base our calculations on the well-known relationship between SNPs and genetic diversity given in (1). From this we develop the concept of an effective sample size, \( n_{\text{eff}} \). This denotes the number of times a query sequence would need to be resequenced, in order to find (on average) the same number of SNPs as can be found in the original alignment.

Implicit in our calculation is the assumption of neutrality. This means that the polymorphisms under consideration have no effect on the fitness of the organism. For any given polymorphism this assumption may not hold. However, since selection mainly affects the distribution of SNPs among sequences rather than the total number of SNPs detected (Tajima, 1989, 1993), our calculations should be rather robust with respect to this concern.

We suggest that the calculation of \( n_{\text{eff}} \) can be used as a standard measure for comparing the number of SNPs detectable from different genomic regions of arbitrary coverage. In particular, when comparing the SNP densities in intergenic and coding sequences the effective sample sizes for both classes of sequence should be taken into account.

ACKNOWLEDGEMENTS

This work was supported by the Max-Planck-Society. We would like to thank D. Altschuler for helpful discussion and anonymous reviewers for their comments on an earlier version of this manuscript. Special thanks to S. Gebauer-Jung for programming assistance.

REFERENCES


