How independent are the appearances of \( n \)-mers in different genomes?

Yuriy Fofanov\(^1\),*, Yi Luo\(^1\), Charles Katili\(^1\), Jim Wang\(^1\), Yuri Belosludtsev\(^3\), Thomas Powdrill\(^3\), Chetan Belapurkar\(^1\), Viacheslav Fofanov\(^1\), Tong-Bin Li\(^1\), Sergey Chumakov\(^1,4\) and B. Montgomery Pettitt\(^1,2\)

\(^1\)Department of Computer Science and \(^2\)Department of Chemistry University of Houston, 4800 Calhoun Road, Houston, TX 77204-3010, USA, \(^3\)Vitruvius Biosciences, The Woodlands, TX, USA and \(^4\)Department of Physics, University of Guadalajara, Guadalajara, Mexico

Received on October 16, 2003; revised on March 9, 2004; accepted on April 1, 2004
Advance Access publication April 15, 2004

ABSTRACT

Motivation: Analysis of statistical properties of DNA sequences is important for evolutionary biology as well as for DNA probe and PCR technologies. These technologies, in turn, can be used for organism identification, which implies applications in the diagnosis of infectious diseases, environmental studies, etc.

Results: We present results of the correlation analysis of distributions of the presence/absence of short nucleotide subsequences of different length (\( n \)-mers, \( n = 5 \) – \( 20 \)) in more than 1500 microbial and virus genomes, together with five genomes of multicellular organisms (including human). We calculate whether a given \( n \)-mer is present or absent (frequency of presence) in a given genome, which is not the usually calculated number of appearances of \( n \)-mers in one or more genomes (frequency of appearance). For organisms that are not close relatives of each other, the presence/absence of different 7–20mers in their genomes are not correlated. For close biological relatives, some correlation of the presence of \( n \)-mers to identify species with relatively short genome sizes (microbial). In such an approach, the shapes of the frequency distributions for particular short subsequences [2–4mers (Nussinov, 1984; Karlin and Ladunga, 1994; Karlin et al., 1997, 1998; Deschavanne et al., 1999; Sandberg et al., 2001)] have been proposed as a measure to decide what microbial genome we are dealing with, based on a given piece of genome or a whole genome.

The above-mentioned papers deal with the case for frequency of appearance when \( n \) is small, such that the total number of \( n \)-mers, \( 4^n \), is smaller than the genome sequence length, \( M \), \( 4^n < M \). It is clear, that distributions of appearance of \( n \)-mers in this range are essentially different from that for random sequences of the same lengths. Here, we calculate whether a given \( n \)-mer is present or absent (frequency of presence) in a given genome that is not the usual calculated number of appearances of \( n \)-mers in one or more genomes (frequency of appearance). We consider the distribution of organisms have become available. Many sequencing projects are progressing but the number of species and variations is so large that comparative genomics is just now beginning to be feasible. A relevant question arises as to whether there is sufficient material to look at them from a statistical viewpoint (Vainrub et al., 2003).

Statistical analysis of the appearance of short subsequences of length \( n \) called motifs or \( n \)-mers in different DNA sequences (see, e.g. Karlin, 2001), from individual genes to full genomes, is of interest in terms of evolutionary biology. In addition, knowledge of the distribution of appearance of \( n \)-mers is necessary for PCR primer (Fislage et al., 1997; Fislage, 1998) and microarray probe design (Southern, 2001). Several attempts (Nussinov, 1984; Karlin and Ladunga, 1994; Karlin et al., 1997; Nakashima et al., 1997, 1998; Deschavanne et al., 1999; Sandberg et al., 2001) have been made to employ the distributions of appearance for \( n \)-mers to identify species with relatively short genome sizes (microbial).

INTRODUCTION

Several hundred sequencing projects have been already completed, and several complete genomes of large multicellular
The calculation of the frequency of presence of \( n \)-mers for \( n > 10 \) in large genome sequences is challenging because of exponential growth of time/memory usage in brute force algorithms. To be able to perform calculations for \( n > 11 \), new algorithms and special data structures have been developed and implemented (Fofanov et al., 2002a,b), see http://bioinfo.uh.edu/publications/ for details.

In this study, we examined the presence/absence of short subsequences in more than one genome simultaneously obtaining a frequency of presence/absence across multiple genomes. This distribution is not related to how many occurrences of an \( n \)-mer are in a particular genomic set. We performed such analyses separately in four different sets of genomes: RNA-based viruses (789 genomic sequences), DNA-based viruses (616 genomic sequences), microorganisms (110 genomes) and human. In each group, the number of simultaneously present 5–18mers were calculated for each pair of genomes. The fourth group contains 24 human chromosomes, for which the numbers of simultaneously present 7–20mers were calculated for each pair of chromosomes.

**RESULTS**

**Frequencies of presence of \( n \)-mers in different genomes**

As the first step of our analysis we have calculated the amount, \( N(n, G) \), of distinct 5–20 long \( n \)-mers present in each of 1500+ considered genomes (\( G \)). The corresponding results for 114 microbial genomes are shown in Figure 1. The value of \( N(n, G) \) depends on two parameters: \( 4^n \)—the total number of all possible \( n \)-mers, and the genome length, \( M \). In Figure 1, we show the frequency of presence of different \( n \)-mers, \( p = N(n, G)/4^n \), as a function of the ratio \( 4^n/M \). Note, that \( 4^n \) grows very fast when \( n \) increases. For short \( n \)-mers, \( n < 7 \), and long sequences, \( M > 4^n \), a kind of ‘saturation’ can be observed, when all or almost all possible \( n \)-mers are present in the sequence. In turn, when \( M \ll 4^n \), only a small part of the total number of \( n \)-mers appears (and, for instance, in microbial genomes most of \( n \)-mers appear only once). The results for different \( M \) and \( n \) form a well-defined pattern. The upper bound of this pattern is given by a simple analytical formula, which can be found under the assumption of the purely random appearance of \( n \)-mers in genomes (see Appendix A for details):

\[
p = \frac{1}{1+\left(\frac{4^n}{M}\right)}. \tag{1}
\]

This upper bound is shown in the figure as a solid line. Similar results for DNA- and RNA-based viruses and multicellular organisms can be found in the supplementary data. It is worth noting that such a pattern for multicellular organisms is located notably below the expected upper bound, which is in agreement with a significant presence of repeated parts in these genomes.
Independence of appearance of \(n\)-mers in genomes

Correlations of presence of \(n\)-mers in different genomes

The principal goal of our research was to find out how independent/correlated the appearances of \(n\)-mers are in different genomes. One of the possible ways to approach this question is by using the well-known multiplication property for the joint probability of the intersection of events, according to which two events \(A\) and \(B\), can be treated as independent if \(p(A \cap B) = p(A)p(B)\).

Consider a simple example based on three different genomes: (1) *Salmonella typhi* (NC_003198), (2) *Mycobacterium tuberculosis* H37Rv (NC_000962) and (3) *Bacillus subtilis* (NC_000964). A complete set of \(n\)-mers would contain \(4^n\) \(n\)-mers, which, for \(n = 12\), is \(4^{12} = 16,777,216\). We use both strands of the complete genome sequences for our calculations. In the text below \(M\) represents the TSL, and \(N(n, G)\) stands for the number of different \(n\)-mers in genome \(G\). In Table 1, we present the number of different 12mers that occur in each of these three genomes together with the corresponding frequency of presence [i.e. the probability of finding randomly picked 12mers in each genome, \(p = N(12, G)/4^n\)].

The number \(N(n, G_1, G_2)\) of \(n\)-mers \((n = 12)\) that appear in each pair of genomes \((G_1, G_2)\) was also computed (Table 2). Based on this, one can compare the probabilities of finding randomly picked 12mers in two genomes simultaneously with the probabilities calculated using the multiplication rule. As shown in Table 2, the actual and calculated (expected) probabilities do not differ greatly from each other. This allows us treating the presence/absence of randomly picked 12mers in these three genomes as independent events.

We calculated the actual and expected probabilities for each pair of genomes in the three groups (1 000 000+ pairs in total).

Table 1. The frequency of presence of 12mers within the three microbial genomes

<table>
<thead>
<tr>
<th>Genomic (G)</th>
<th>Genome length</th>
<th>Total sequence length (bp)</th>
<th>Number of different 12mers present in genome: (N(n,G))</th>
<th>(p = N(12,G)/4^n) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.typhi</td>
<td>4 809 037</td>
<td>9 618 074</td>
<td>5 813 330</td>
<td>34.65</td>
</tr>
<tr>
<td>M.tuberculosis</td>
<td>4 411 529</td>
<td>8 823 058</td>
<td>4 361 508</td>
<td>26.00</td>
</tr>
<tr>
<td>B.subtilis</td>
<td>4 214 814</td>
<td>8 429 628</td>
<td>5 346 103</td>
<td>31.87</td>
</tr>
</tbody>
</table>

Table 2. Actual and predicted simultaneous presence of 12mers within the three microbial genomes: (1) *S.typhi*, (2) *M.tuberculosis* H37Rv and (3) *B.subtilis*

<table>
<thead>
<tr>
<th>Case</th>
<th>Number of 12mers</th>
<th>(N(n, G_1, G_2)/4^n)</th>
<th>Calculated probability assuming independence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present in genomes (1) and (2)</td>
<td>1 943 814</td>
<td>11.6</td>
<td>9.0</td>
</tr>
<tr>
<td>Present in genomes (1) and (3)</td>
<td>2 335 710</td>
<td>13.9</td>
<td>11.0</td>
</tr>
<tr>
<td>Present in genomes (2) and (3)</td>
<td>1 334 288</td>
<td>8.0</td>
<td>8.3</td>
</tr>
</tbody>
</table>

We were especially interested in the range of \(n\) which gives rise to the frequency of presence, \(p^*\), of different \(n\)-mers in the genome between 5% and 50% of the total possible number of possible \(n\)-mers \((4^n)\). This range for different microbial
Table 3. The optimal length of n-mers ($n^*$) for different genome sizes and frequencies of presence ($p^*$)

<table>
<thead>
<tr>
<th>Total sequence length (bp) (Mb)</th>
<th>$n^<em>$ determined for frequency of presence 50% ($p^</em> = 0.5$)</th>
<th>$n^<em>$ determined for frequency of presence 5% ($p^</em> = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>9.80</td>
<td>11.93</td>
</tr>
<tr>
<td>2.0</td>
<td>10.47</td>
<td>12.59</td>
</tr>
<tr>
<td>10.0</td>
<td>11.63</td>
<td>13.75</td>
</tr>
</tbody>
</table>

genome sizes can be numerically determined from Figure 1. The corresponding frequency of presence for purely random sequences (random boundary) is shown in Figure 1 by a solid line. The analytical formula for the random boundary can be used to estimate this range analytically:

$$n^* = \frac{\log[M(1 - p^*)/p^*]}{\log(4)}. \quad (2)$$

This formula works well for all the three groups of genomes (viruses, microbes and multicellular organisms). The upper and lower bounds of $n^*$ for genome sizes between 0.8 and 10 Mb, which are typical for microbialis, are shown in Table 3. In accordance with this, the value $n = 12$ seems to be the most reasonable one for all microbial genomes. For viral genomes the appropriate value was found to be $n = 7$.

We found that for all 11,990 pairs of microbial genomes and the value of $n = 12$ the average ratio of actual and expected probabilities is $1.37 \pm 0.67$. For viral genomes and the value of $n = 7$, the average ratio of actual and expected probabilities was found to be $1.07 \pm 0.12$ for 387,840 genome pairs DNA-based viruses and $1.10 \pm 0.12$ for 621,732 genome pairs RNA-based viruses. Thus, we conclude that for this range of $n$ the presence of $n$-mers in different genomes, to a good approximation, can be treated as independent events.

The highest deviations between the expected and actual probabilities were found among closely related genomes. For instance, using 7mers, a high ratio (185%) was found for Duck hepatitis B virus (NC_001344) versus Stork hepatitis B virus (NC_003325) with 8.1% expected and 15.0% actual.

An example of closely related microbial genomes would be *Staphylococcus aureus* N315 (NC_002758) versus *Staphylococcus Mu50* (NC_002758) with 4.0% expected and 19.7% actual. The highest deviations between the expected and actual probabilities was found among closely related genomes. For viral genomes and the value $n = 7$, the average ratio of actual and expected probabilities was found to be $1.07 \pm 0.12$ for 387,840 genome pairs DNA-based viruses and $1.10 \pm 0.12$ for 621,732 genome pairs RNA-based viruses. Thus, we conclude that for this range of $n$ the presence of $n$-mers in different genomes, to a good approximation, can be treated as independent events.

We performed the same calculation for the 24 human chromosomes pairwise. The average ratio of actual and expected probabilities of 14mers is $1.91 \pm 0.16$, maximum ratio being found for 20th and Y-chromosomes (expectation 2.9% versus actual 6.9%).

**DISCUSSION**

**Microbial/viral fingerprints using random subsets of n-mers**

It may be assumed that our results for 1500+ genomes can be extended to other genomes (many yet to be sequenced). In this case one may use relatively small sets of randomly picked $n$-mers for differentiating between different viruses and organisms.

This idea can be illustrated by continuing our example for three microbial genomes. Let $n^*$ be the size of $n$-mer, which fits the interval where from 5% to 50% of all possible $n$-mers show up for a desirable range of genome lengths. In accordance with Table 3, we may choose the value $n^* = 12$. Let us randomly pick $L$ 12mers (say, $L = 1000$). For example, $L$ can be the number of probes placed on a microarray. Given a genome $G_1$ with the frequency of presence of $n$-mers $p_1$, we expect that $K = p_1L$ n-mers present in $G_1$ will appear also in our random set, forming a ‘fingerprint’ of $G_1$ (in our example, we expect $50 < K < 500$). The probability, $\epsilon$, that the fingerprint of $G_1$ will exactly coincide with the fingerprint of another genome $G_2$ will be $\epsilon = (\frac{p_2}{p_1})^{K}$. The value of $\epsilon$ can be used as a basis of an algorithm for genome fingerprinting.
of some other genome $G_2$ (with the frequency of presence of $n$-mers $p_2$) is found in Appendix B. The result is

$$\epsilon = (1 - p_1 - p_2 + 2p_{12})^L.$$

(3)

Here $p_{12}$ is the probability for the $n$-mer to be present in both genomes simultaneously. Let us consider the numeric example from Tables 1 and 2 of two species that are far from each other, $S.\text{typhi}$ versus $M.\text{tuberculosis}$ H37Rv; $p_1 = 0.3465$, $p_2 = 0.2600$, $p_{12} = 0.1160$. With $L = 1000$, a remarkable accuracy of $\epsilon = 1.7 \times 10^{-304}$ can theoretically be achieved.

Given a desirable probability of error, $\epsilon$, one can determine the appropriate size, $L$, of a random set of $n$-mers which can be used for reliable identification of genomes as

$$L = \frac{\log \epsilon}{\log(1 - p_1 - p_2 + 2p_{12})}. \quad (4)$$

For related organisms, the genomes may contain large common parts. This means that $p_{12}$ may be close to $p_1$ and $p_2$. To give a numeric example of close relatives, let us consider $S.\text{aureus}$ N315 versus $S.\text{aureus}$ Mu50. Now $p_1 = 0.198$, $p_2 = 0.203$, $p_{12} = 0.197$ and an accuracy of $\epsilon = 10^{-10}$ can be achieved with $L = 3278$. We would like to stress the logarithmic dependence of the sampling size on the accuracy, $\epsilon$. This feature is of principal importance for our discussion.

Therefore, we can use practically any sufficiently random subset of $n$-mers of appropriate size to construct a microarray to diagnose to which organism a given DNA/RNA sample belongs. Different sizes of $n$-mers must be employed for recognition of different organisms based on their genome lengths. Values of $n$ that correspond to given intervals of genome lengths can be easily calculated using above formulas. In fact, only 11 different $n$ values, $7 \leq n \leq 17$, would be enough to cover a large variety of genome sizes from 1 kb to 9 Gb.

The important advantage of such an approach is that it can be used without a priori knowledge of the sequence itself. This implies that there is no need to perform the expensive and time-consuming process of sequencing before array design. It is enough to obtain the purified DNA, hybridize it on a sufficiently random microarray chip and check which $n$-mers show up. Taking into account how accessible the DNA of thousands of microbial and viruses are, how easily each microarray can be produced, and the fact that we do not need to determine quantitative values of expression (we need just a yes/no answer)—it should be possible to produce an essentially universal microbial/viral DNA chip.

Fingerprints of closely related organisms

We next consider what happens when we try to compare closely related organisms using this approach (e.g. different types of influenza or different strands of the same microbes).

We assume that two genomes $G_1$ and $G_2$ almost coincide and differ only in $m$ randomly located nucleotides. This situation simulates the existence of point mutations or single nucleotide polymorphisms (SNPs). Let $L$ be the size (number of probes) of the microarray and $p$, the frequency of presence of $n$-mers in a genome with a TSL value $M$. The value of $L$, necessary to distinguish the fingerprints of two genomes with the error probability $\epsilon$, can be estimated by the formula (see Appendix B):

$$L = \frac{\log \epsilon}{p \log(1-nm/N)} \leq \frac{M|\log \epsilon|}{pmn}. \quad (5)$$

Here, $N$ is the number of different $n$-mers contained in $G_1$ (which is approximately equal to the number of different $n$-mers contained in $G_2$).

Such an approach can provide the level of accuracy necessary for the individual human fingerprints. Let us assume that the differences between individual human beings appear only because of SNPs, which have equal probability and are randomly located in genome. According to literature estimates (Weiner and Hudson, 2002), the total number of SNPs in the human genome is expected to be $\sim 3000000$. Then, calculating the necessary size for the random microarray $(m/M \sim 0.1\%, \epsilon = 10^{-10}, n = 17, p = 0.284)$ we have $L \sim 4769$. This rough estimation is promising and indicates that this possibility deserves a proper experimental study.

We would like to recall, that our theoretical estimations have been made for randomly picked sets of $n$-mers. The further possibility exists to start with a larger than necessary random set of $n$-mers (say, $L = 10000$) and then to decrease the microarray size experimenting with the desirable set of genomes (using, for instance, a simple optimization approach).

Conclusions

We presented results of a correlation analysis for distributions of the presence/absence of short subsequences of different length ($n$-mers, $n = 5 - 20$) in more than 1500 microbial and viral genomes together with five genomes of multicellular organisms (including human). Our results show that for organisms that are not close relatives to each other, a range of values of $n$ can be found, such that the presence/absence of different $n$-mers in different genomes are practically not correlated (within a probabilistic tolerance, $\epsilon$). For close relatives such correlations do appear, but are not as strong as might be expected.

The size of the correlations among the $n$-mers present in different genomes lead to the possibility of using random sets of $n$-mers (with appropriately chosen $n$) to discriminate between different microbial and viral genomes, and possibly, individual human beings with a convenient number of combinatorial experiments. The formulas derived, yield the size of a combinatorial experiment designed to identify an organism given the length of its genome, a convenient length of probe, $n$ and a tolerance or error, $\epsilon$.
Clearly, additional experimental study (including, e.g. hybridization of microbial samples on random microarrays) is necessary to verify if the statistical features described above can lead to the creation of a real biosensor (as it is suggested by our in silico experiments). Future studies should take into account errors in the course of hybridization. Rough theoretical estimation, assuming independent probabilities of hybridization error at different microarray sites, suggests that total number of hybridization errors on the array of size \( L \) is proportional to \( \sqrt{L} \). Thus, the total relative error due to imperfect hybridization can be made small by increasing the number of probes on the microarray \( L \). On the other hand, it is currently not clear to what degree the genomes are correctly assembled. Possible errors in sequences may have affected our results. We however believe that the parts of genomic sequences that have been correctly reconstructed are significant enough to determine the statistical properties described above.

**ACKNOWLEDGEMENTS**

The authors thank Prof. M. Hogan for interesting conversations. S.C., B.M.P. and Y.F. thank TLCC for partial funding of this work. T.-B.L. was supported by a training fellowship from the W.M. Keck Foundation to the Gulf Coast Consortium through the Keck Center for Computational and Structural Biology. B.M.P. and Y.F. thank the NIH for partial support of this work and NPACI for computational support. S.C. is grateful to the University of Houston Computer Science Department for hospitality.

**REFERENCES**


Fofanov,Y., Fofanov,V. and Pettitt,B.M. (2002a) Fast subsequence search using incomplete search trees. The Seventh Structural Biology Symposium of Sealy Center for Structural Biology, The University of Texas Medical Branch, Galveston, TX, p. 51.


**APPENDIX A**

Here, we will analytically estimate the frequency of presence of \( n \)-mers in a genome of length \( M \). Let us apply the logic of the example shown in Tables 1 and 3 to autocorrelations, i.e. let us check whether the appearances of distinct \( n \)-mers are independent or correlated within a single genome. Assume that the multiple appearances of a given \( n \)-mer at different locations within the same genome are also independent events. Then, the probability of \( n \)-mer to appear once is \( p \), twice is \( p^2 \), thrice is \( p^3 \) and so on. The total number of \( n \)-mers in the genome, taking into account multiple appearances is

\[
M \approx 4^n (p + p^2 + p^3 + \cdots) = \frac{4^n p}{1 - p},
\]

(A1)

from which one obtains

\[
p \approx \frac{M}{(M + 4^n)}. \quad \text{(A2)}
\]

This formula has been presented in the text, and is shown in Figure 1 by a solid line. One may also compare it with the experimental values from the last column of Table 1. In accordance with Equation (1) we have for *S.typhi* \( p = 34.44\% \), for *M.tuberculosis* H37Rv, \( p = 34.46\% \) and for *B.subtilis* \( p = 33.44\% \). This corresponds better to experimental values (34.65, 26.00 and 31.87%, respectively)
than the estimation without taking into account multiple appearances,

$$p \approx \frac{M}{4^n}, \quad \text{(A3)}$$

which leads to the probabilities 57.3, 52.6 and 50.2%, respectively. This fact is in accordance with the conclusion about the apparently nearly random statistical character of the appearance of \(n\)-mers in a single genome.

**APPENDIX B**

Here, we will estimate the probability to make an error distinguishing organisms by their analysis (‘fingerprints’) in a random microarray, which consists of \(L\) \(n\)-mers. Assume that we need to discriminate between the two genomes \(G_1\) and \(G_2\) of sizes \(M_1\) and \(M_2\), respectively. Let \(G_1 (G_2)\) contains \(N_1 (N_2)\) different \(n\)-mers and \(N_{12} = N(n, G_1, G_2)\) \(n\)-mers are present simultaneously in both genomes (this is the size of intersection of two sets of \(n\)-mers corresponding to ‘\(n\)-mer contents’ of \(G_1\) and \(G_2\); we denote this set as \(G_1 \cap G_2\)). The union \(G_1 \cup G_2\) contains \(N_1 + N_2 - N_{12}\) \(n\)-mers. Let us consider a fingerprint of the union of the two genomes, \(G_1 \cup G_2\). For every \(n\)-mer appearing in this fingerprint, the probability that it occurs in the intersection region, \(G_1 \cap G_2\), is

$$\frac{N_{12}}{N_1 + N_2 - N_{12}}. \quad \text{(A4)}$$

An error, \(E\), occurs when two genomes share the same fingerprint, i.e. all \(n\)-mers that form the fingerprint represent the intersection region. This will happen with probability

$$P(E | k) = \left(\frac{N_{12}}{N_1 + N_2 - N_{12}}\right)^k. \quad \text{(A5)}$$

In fact, this is a conditional probability of an error, \(E\), if we have a fingerprint of length \(k\).

We now need to calculate an average with respect to all possible fingerprints. There are \(C_k^L = L!/[k!(L-k)!]\) different fingerprints of the size \(k\), which appear with equal probabilities \([P(S \in G_1 \cup G_2)]^k [1 - P(S \in G_1 \cup G_2)]^{L-k}\), where \(P(S \in G_1 \cup G_2)\) is the probability for \(n\)-mer \(S\) to find itself in the intersection \(G_1 \cup G_2\) sampling \(L\) times. Therefore, we come to a binomial distribution of fingerprint sizes,

$$P(k) = \frac{L!}{k!(L-k)!} \left[ \frac{N_1 + N_2 - N_{12}}{4^n} \right]^k \times \left[ 1 - \frac{N_1 + N_2 - N_{12}}{4^n} \right]^{L-k}. \quad \text{(A6)}$$

Calculating the average error we have,

$$P(E) = \sum_k P(E | k) P(k) = (1 - p_1 - p_2 + 2p_{12})^L. \quad \text{(A8)}$$

Here, \(p_j = N_j/4^n\) is the probability of presence in \(G_j\) (\(j = 1, 2\)), and \(p_{12} = N_{12}/4^n\) is the probability of presence in the intersection \(G_1 \cap G_2\). Given a desirable level of tolerance or error, \(P(E) \sim \varepsilon\), one can now estimate the appropriate combinatorial experiment (array) size:

$$L = \frac{\log \varepsilon}{\log(1 - p_1 - p_2 + 2p_{12})}. \quad \text{(A9)}$$

We would like to again stress the logarithmic dependence of the microarray size \(L\) on the error level \(\varepsilon\). This feature is of principal importance for the analysis under discussion. The following three cases will be considered separately.

**Essentially different organisms**

In this case, in accordance with the discussion in the text, the presence/absence of \(n\)-mers in one genome is not correlated with the presence/absence of \(n\)-mers in another genome and we can write \(p_{12} \approx p_1 p_2\). Taking, for simplicity, \(p_1 \approx p_2 \approx p\), we obtain,

$$L = \frac{\log \varepsilon}{\log(1 - 2p + 2p^2)}. \quad \text{(A10)}$$

For instance, if \(\varepsilon = 10^{-10}\) and \(p = 0.05\), we obtain \(L = 230\).

**Related organisms**

Now, \(p_{12} \neq p_1 p_2\). Assuming that intersection \(G_1 \cap G_2\) almost coincides with the union, \(G_1 \cup G_2\), or

$$N_1 + N_1 - N_{12} > N_{12} \gg N_1 + N_2 - 2N_{12}, \quad \text{(A11)}$$

one can rewrite Equation (A9) in a slightly different form. Starting once again with Equations A7–A9 and approximating the binomial distribution by the Gaussian of width \(s = \sqrt{LP(1-P)}\), centered at \(k = LP\) where \(P = (N_1 + N_2 - N_{12})/4^n\) is the probability for an \(n\)-mer to be present in the union \(G_1 \cup G_2\) we find,

$$P(E) = \sum_k e^{-ak} \frac{1}{s \sqrt{2\pi}} e^{-(k-\bar{k})^2/2s^2},$$

$$e^{-a} = \frac{N_{12}}{N_1 + N_2 - N_{12}}. \quad \text{(A12)}$$

Provided that \(a \ll 1\) [which follows from inequality (5)] and \(\bar{k} \gg 1\) (which is consistent with a small error level), one can change the summation to integration and obtain immediately,

$$P(E) = \frac{1}{s \sqrt{2\pi}} \int e^{-a(k-\bar{k})^2/2s^2} \, dk = e^{-a \bar{k}^2/2s^2}. \quad \text{(A13)}$$

Finally,

$$P(E) \approx \frac{N_{12}}{N_1 + N_2 - N_{12}}. \quad \text{(A14)}$$

Now we can find the relation between the error level and the microarray size in the form,

$$\bar{k} = PL = \frac{\log \varepsilon}{\log[N_{12}/(N_1 + N_2 - N_{12})]}. \quad \text{(A15)}$$

Here, \(P\), the probability of presence of \(n\)-mer in the intersection of two genomes, is given by
\[ P = (N_1 + N_2 - N_{12})/4^n \approx p_1 \approx p_2. \] The last formula leads to similar numerical values as Equation (A1) if \( N_{12} \gg N_1 + N_1 - 2N_{12} \). Say, for \( P = 0.05 \), \( N_{12}/(N_1 + N_2 - N_{12}) = 0.9 \), \( \epsilon = 10^{-10} \), we have, \( L = 4371 \).

**Closely related organisms**

Let us assume that two genomes \( G_1 \) and \( G_2 \) almost coincide and differ only in \( m \) randomly located characters (nucleotides). This situation simulates the existence of SNPs. For simplicity, let us assume, that \( N_1 = N_2 = N \). Every character that is different in \( G_1 \) and \( G_2 \) belongs simultaneously to \( n \) different \( n \)-mers, and the size of the subset in \( G_1 \cup G_2 \) which consists of the \( n \)-mers that are different in \( G_1 \) and \( G_2 \) has a size, \( nm = 2N - 2N_{12} \). Then,

\[ N_{12} = N - \frac{mn}{2}, \quad \text{or} \quad N_1 + N_2 - N_{12} = N + \frac{mn}{2}. \]

Taking into account, that \( N \leq M \), we arrive at the estimation,

\[ P(E) \approx \left(1 - \frac{nm}{N}\right)^{\frac{L}{n}} = \epsilon. \]  

(A16)

\[ L = \frac{\log \epsilon}{P(n)} \leq \frac{M |\log \epsilon|}{Pmn}. \]  

(A17)