Published in final edited form as:
J Biol Phys Chem. 2005 December 1; 5(4): 121–128.

The theoretical basis of universal identification systems for bacteria and viruses

S. Chumakov[1], **C. Belapurkar**[1], **C. Putonti**[1,*], **T.-B. Li**[1], **B.M. Pettitt**[1,2,3], **G.E. Fox**[2,4], **R.C. Willson**[2,4], and **Yu. Fofanov**[1,2]

[1] Department of Computer Science, University of Houston, Houston, TX 77204, USA

[2] Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA

[3] Department of Chemistry, University of Houston, Houston, TX 77204, USA

[4] Department of Chemical Engineering, University of Houston, Houston, TX 77204, USA

Abstract

It is shown that the presence/absence pattern of 1000 random oligomers of length 12–13 in a bacterial genome is sufficiently characteristic to readily and unambiguously distinguish any known bacterial genome from any other. Even genomes of extremely closely-related organisms, such as strains of the same species, can be thus distinguished. One evident way to implement this approach in a practical assay is with hybridization arrays. It is envisioned that a single universal array can be readily designed that would allow identification of any bacterium that appears in a database of known patterns. We performed *in silico* experiments to test this idea. Calculations utilizing 105 publicly-available completely-sequenced microbial genomes allowed us to determine appropriate values of the test oligonucleotide length, *n*, and the number of probe sequences. Randomly chosen *n*-mers with a constant G + C content were used to form an *in silico* array and verify (a) how many *n*-mers from each genome would hybridize on this chip, and (b) how different the fingerprints of different genomes would be. With the appropriate choice of random oligomer length, the same approach can also be used to identify viral or eukaryotic genomes.

Keywords

microbe identification; oligonucleotide microarray fingerprinting; species identification

1. INTRODUCTION

All species of living organisms have unique genomes whose nucleotide sequence distinguishes them from any other organism. Molecular approaches for identifying bacteria and viruses have relied on identifying one or more unique subsequences (oligonucleotides, or "*n*-mers") in these genomes, whose presence in a sample can be used as an individual microbial/viral "fingerprint" [1–4]. There are a variety of ways in which such a unique subsequence can be used to devise specific assays. For example, the appropriate oligomers can be used as PCR primers or as hybridization probes [5]. This approach is limited by the fact that it depends on specific knowledge about a target sequence within the organism of interest. If the organism is a novel isolate this information will not be available. If it is a close relative of a known strain, then the target may not be sufficiently unique to distinguish the two. This can be a major practical obstacle. For example, a probe targeting the 16S rRNA gene of *Bacillus anthracis*

*Corresponding author. putonti@bioinfo.uh.edu.

would be likely to detect non-pathogenic *B. thuringensis* strains as well. Moreover, any single unique subsequence found by analysing known genomes or specific genes may always be present by chance in some other genome. Therefore, what is thought to be a unique identifier may ultimately fail at an inopportune time. Finally, the unique identifier can be present in an organism under study but nevertheless not available for hybridization for experimental reasons such as cleavage, inaccessibility, etc.

An alternative solution to identification utilizes sets of the sequences (probe sets) selected from a reference sequence. Such a probe set can be designed to encompass a known set of polymorphic sequences for a specific portion of the genomic DNA. For instance, Gingeras et al. [6] designed a set of probes capturing known polymorphisms within the *rpoB* gene sequence for *Mycobacterium tuberculosis*. The resulting high-density oligonucleotide array was able to uniquely identify *Mycobacterium* isolates. It can however be inferred from the literature [6] that the array would not be able to reliably and successfully identify species that do not share homologous regions with the *rpoB* gene of *M. tuberculosis*. Probe sets have also been designed using the *Escherichia coli* genomic sequence as a reference, specifically with microbe identification in mind [7,8]. In both cases cited, subsequences known to occur frequently in *E. coli* were selected for probes. This probe set was then used for identification of closely-related species (*Xanthomonas* [7] and *Salmonella enterica* isolates [8]). When considering more distantly-related organisms or more complex ones, e.g., multicellular organisms, *E. coli* may not serve as an appropriate reference sequence.

Herein we present the theoretical underpinnings for an alternative strategy suggested by genome sequence comparisons. The essence of the idea is to collect experimental information about the presence or absence of each member of a set of non-specific (even randomly chosen) short subsequences (*n*-mers) in a genome of interest. The total number of possible oligonucleotides, 4^n, increases as oligonucleotide length *n* increases. For sufficiently small values of *n*, a given genome will be large enough for all possible *n*-mers to be present. For greater values of *n*, the number of alternative *n*-mer sequences becomes extremely large and despite the considerable size of a genome, most *n*-mers will never occur. For a given genome, however, there is always an intermediate value of *n*, for which a reasonable fraction of oligomers will occur, while many do not. If enough probes of these intermediate lengths are monitored, the pattern of presence and absence should constitute a highly unique “molecular fingerprint” of the organism being characterized. This pattern can be readily matched to a library of known patterns to identify a bacterial or viral strain, or to identify related organisms.

One way to match this approach to experimental practice is to print probes complementary to a representative number of *n*-mers on a single cDNA array. When DNAs from different microbes are hybridized to the arrays, patterns will be produced indicating the presence/absence of each *n*-mer in each genome. We show herein that with a reasonable array size (~1000 spots) and appropriate lengths of the subsequences ($n = 12–13$ nucleotides for bacteria) the probability of error (the probability that two different organisms will have the same pattern) is negligibly small [9,10]. Therefore, such a random array can be used as a “universal identifier”.

The work presented here establishes the theoretical feasibility of the approach by showing that more than enough probes exist to construct an experimentally reasonable universal identifier array. *In silico* hybridizations using such arrays are used to examine the utility of the arrays in distinguishing both distantly- and closely-related organisms. Analysis of these results allows us to determine the best values of *n* and *m* (where *m* is the number of G and C nucleotides in the *n*-mer—its GC content). The results obtained clearly indicate that a universal random array may be used for efficient identification of microbial organisms (including close relatives). Original data and results can be found on the supplementary data website http://www.bioinfo.uh.edu/publications/universal_identification/.

2. METHODS

2.1. Computational background

The sequences of 105 complete microbial genomes with sizes ranging from 0.58 Mb to 9.11 Mb were obtained from the NCBI database. Performing a statistical analysis of long subsequences (of sizes ≥ 11) is a challenging task. Unique algorithms and specialized data structures (counting arrays and incomplete search trees) were developed for this purpose [11–13]. These algorithms provided superior time and memory efficiency for the computations.

2.2. Hypothesis of randomness

The total number of *n*-mers in a genome *G* is approximately equal to the genome length *M*. Let N_G be the number of different *n*-mers in *G*. We introduce the frequency of the presence of *n*-mer *S* in a genome *G* as $f(S,G) = N_G/4^n$. Note here that the frequency of presence is not the same as the frequency (or probability) of occurrence within a given genome (where the actual number of occurrences is also taken into account); see Appendices A and B for more details. We have recently shown by analysing a large number of publicly available sequenced genomes [10] that in the range $M_G < 4^n$, the frequency of presence is very close to that theoretically expected in a genome of random sequence. For the case of equal single-nucleotide probabilities, $p_A = p_C = p_G = p_T = 1/4$, we have for a random genome,

$$f_0 = 1 - \exp\left(-\frac{M}{4^n}\right). \tag{1}$$

This frequency is the same for all *n*-mers with a given *n*. When $M \ll 4^n$, we have $f_0 \sim M/4^n$. The existence of a range where the *n*-mer content of genomes can be considered as nearly random (a “random domain”) provides a starting point for our approach.

In order to take the GC content into account, one may generalize eqn 1, allowing for the different single nucleotide probabilities. Hybridization rules dictate that in the overall two-strand genome, the number of A nucleotides will be equal to the number of T, and a similar correspondence will be found between G and C. We have also found that analysis of a single strand sequence (see Figure 1) will usually give a similar fraction of A as of T (and G as of C) on a single strand, i.e. there is no strand preference for particular nucleotides. Therefore, we accept that $p_G = p_C = p_1$ and $p_A = p_T = 1/2 - p_1$.

However, nucleotides A and C can appear with different probabilities: $p_1 \neq 1/2 - p_1$. Let n_d be the number of nucleotides *d* (= A, C, G, T) in a given *n*-mer, such that $n_G + n_C + n_A + n_T = n$. The frequency of the presence of *n*-mer *S* with the GC content $m = n_G + n_C$ in a random genome takes on the form (see appendix A),

$$f(S) = 1 - e^{-Mp_S},$$
$$p_S = p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G} = p_1^m \left(\tfrac{1}{2} - p_1\right)^{n-m}. \tag{2}$$

Below we will compare this equation, which is valid for an idealized random genome, with the results of our computational experiments on real microbial genomes.

2.3. Choosing the value of *n*

Let us consider a typical hybridization experiment. In accordance with the random genome model, eqn 2, we expect that out of L n-mers placed on a chip, Lf n-mers will be present in the genome G. Let G be exposed to the chip, and the n-mer S be present both in the genome and in the set of probes placed on the chip. Since our experiment is not perfect, S will not necessarily be detectable on the chip [10]. Assume that the probability of detected hybridization is the same for any probe and equals p_{hyb} (for an ideal experiment, $p_{hyb} = 1$). The number of probes expected to hybridize (the average size of a fingerprint), K, is given by:

$$\langle K\rangle = p_{hyp} Lf.$$

Assuming that hybridization at different array sites precedes independently, the uncertainty in the fingerprint size can be found using the binomial distribution,

$$\sigma_K = \sqrt{p_{hyb}\left(1 - p_{hyb}\right) Lf}.$$

We wish the relative uncertainty in the fingerprint size to be small,

$$\frac{\sigma_K}{\langle K\rangle} = \sqrt{\frac{p_{hyb}}{Lf\left(1 - p_{hyb}\right)}} \sim \frac{1}{\sqrt{Lf}} \ll 1$$

e.g. choosing the relative uncertainty ~ 0.3, we have $Lf \sim 10$. Therefore, f cannot be extremely small, for in such a case the array size would have to be excessively large to provide reliable identification. In accordance with eqns 1 and 2, the frequency of presence depends on two parameters: the genome size M and the oligonucleotide size n. M is set by the type of organism of interest. For bacteria, the genome size M is in the range $10^6 < M < 10^7$, see Figure 2. However, one can achieve a manageable value of f by changing n. The prediction of formula (1), in agreement with experimental experience for microbial genomes, leads to the choice $n = 12$, which ensures that f is in the range $0.1 < f < 0.5$. This estimation from the random genome model is in good agreement with the results of computer experiments with real genomes.

2.4. Probability of error in identification of microbial genomes

Let us estimate the probability of error in discriminating organisms by their fingerprints on a random array which consists of L n-mers. Assume that we need to discriminate between two genomes G_1 and G_2 of corresponding sizes M_1 and M_2. Let G_1 contain N_1 and G_2 contain N_2 different n-mers, and let N_{12} n-mers be present simultaneously in both genomes (this is the size of the intersection $G_1 \cap G_2$ of the n-mer content of G_1 and G_2). The probability ε that the random array fails to distinguish the two genomes is equal to $\varepsilon = [1-(N_1 + N_2 - 2N_{12})/4^n]^L$ (see Appendix C for the derivation of ε). Given an acceptable error probability ε one can now estimate the appropriate array size:

$$L = \frac{\log\varepsilon}{\log\left[1 - (N_1 + N_2 - 2N_{12})/4^n\right]}. \quad (3)$$

We would like to stress the logarithmic dependence of the array size L on the error level ε (or, equivalently, the exponential decay of the error level with increasing array size). This feature is of principal importance for the analysis under discussion.

3. RESULTS AND DISCUSSION

We generated *in silico* random 12-mer arrays of size $L = 1000$ with various values of GC content m. The expected fingerprint that would be seen on the random array (the set of n-mers present simultaneously in the genome and on a given chip) was generated for each of 105 real, sequenced bacterial genomes. Typical results are shown in Figures 3–7. The sizes of the microbial fingerprints on chips with GC content $m = 6$ and $m = 4$ are plotted in Figures 3 and 4 as the red curves. The fingerprint sizes for random "genomes" with the same lengths and GC content, calculated from eqn 2, are shown by the blue curves. The results of the random genome model with equal nucleotide probabilities, eqn 1, are plotted as green curves. Every point corresponds to a genome; the horizontal axis is the genome length. An agreement is observed between the real genome fingerprints and ones generated from random "genomes". This confirms our hypothesis of the random presence of n-mers in genomes in the size range $M < 4^n$.

Next we considered the intersections of fingerprints (number of common n-mers) for every genome pair ($105 \times 104/2 = 5460$ pairs in total). The results for the arrays with GC content $m = 6$ and $m = 8$ are depicted in Figures 5 and 6 as scatter plots of intersection sizes for real genomes against the intersection sizes for random-sequence "genomes" of the same length and GC content (in accordance with eqn 2). For species that are not close relatives of each other, we have a high correlation of the results for real and random genomes, i.e. the points on the scatter plot are close to the solid ($x = y$) line. Intersection sizes for close relatives are larger than those arising randomly, and the corresponding points appear above the $x = y$ line (see also [9]). In Figures 5–7 we label the genomes that have the most similar fingerprints; all of which are closely related. These are four strains of *Escherichia coli* and *Shigella flexneri*, two strains of *Mycobacterium tuberculosis*, two strains of *Streptococcus pneumoniae*, etc.; see Table 1.

Even for close relatives the difference in fingerprints is sufficient to distinguish between them. This is demonstrated in Figure 7, where we show the distribution of ratios of the size of the intersection of the two fingerprints, $F_1 \cap F_2$, over the size of the union of the same two fingerprints, $F_1 \cup F_2$,

$$R = \frac{\text{size}(F_1 \cap F_2)}{\text{size}(F_1 \cup F_2)}$$

(for identical fingerprints, $R = 1$). Again, the real genome results are shown against the corresponding results from random-sequence "genomes" with the same lengths and GC contents. The principal part of the graph is formed by the species that are not close relatives of each other. Far above lie the points representing close relatives (see Table 1). Note, that the fingerprints of all close relatives still differ in several probes (n-mers), and, generally, can be discriminated. This illustrates that substantial differences occur even at strain level, and even modest-sized arrays (by today's standards) will be able to distinguish very closely related strains. Higher resolution can be achieved with larger array sizes.

4. CONCLUSIONS

We have compared the fingerprint sizes for different microbes and the sizes of intersections of fingerprints for all of the microbial pairs, with the corresponding results for random-

sequence genomes of the same length and GC content. Remarkably good correspondence with the random model has been found. This means that the *n*-mer size and GC content can be chosen in such a way that all of the microbes have fingerprint sizes big enough to be identified and the fingerprints of different microbes are sufficiently different to ensure a very low probability of misidentification. Fingerprint sizes, intersection sizes and the probability of misidentification can be estimated from the random genome model.

The results of the *in silico* experiments presented in this report indicate that discrimination of bacterial species is achievable with randomly designed arrays. Furthermore, the random-array approach provides distinguishable fingerprints even for closely related organisms. Thus, the probability of misidentification as a result of subsequent genomic mutation is negligible due to the fact that the pattern has a one-to-one or linear correspondence to the difference between the sets of *n*-mers in each genome. It is also clear that the same approach will work with viruses and eukaryotic organisms, though the numerical parameters will change.

The overriding advantage of the random-array approach described here is that a single experimental system can be used for any and all bacteria. No sequencing will be required. One will only need to compare the pattern appearing on the array against a library of known patterns. In addition to the constraints described here (on GC content), the arrays can be further constrained to not hybridize to known backgrounds of various types, e.g., human and human SNP DNA. Finally, it should be pointed out that the essence of the method is a cumulative summary of yes/no distinctions for a large number of probes. Hence alternative versions of the assay can be constructed in which the pattern is generated computationally after the data is collected. For example, a series of PCR amplifications might be used to generate the yes/no distinctions.

Acknowledgments

This work was partially supported by grants from the Texas Learning and Computational Center (to YF, BMP, RCW and GEF). TBL's work was supported in part by the Keck Center for Computational and Structural Biology. CP's work was supported by a training fellowship from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NLM Grant No. 5T15LM07093).

References

1. Barrangou R, Yoon SS, Breidt F Jr, Fleming HP, Klaenhammer TR. Characterization of six *Leuconostoc fallax* bacteriophages isolated from an industrial sauerkraut fermentation. Appl Environ Microbiol 2002;68:5452–5458. [PubMed: 12406737]
2. Genco RJ, Loos BG. The use of genomic DNA fingerprinting in studies of the epidemiology of bacteria in periodontitis. J Clin Periodontol 1991;18:396–405. [PubMed: 1890219]
3. Nguimbi E, Li YZ, Gao BL, Li ZF, Wang B, Wu ZH, Yan BX, Qu YB, Gao PJ. 16S–23S ribosomal DNA intergenic spacer regions in cellulolytic myxobacteria and differentiation of closely related strains. Syst Appl Microbiol 2003;26:262–268. [PubMed: 12866853]
4. Seurinck S, Verstraete W, Siciliano SD. Use of 16S–23S rRNA intergenic spacer region PCR and repetitive extragenic palindromic PCR analyses of *Escherichia coli* isolates to identify nonpoint fecal sources. Appl Environ Microbiol 2003;69:4942–4950. [PubMed: 12902290]
5. Reyes-Lopez MA, Mendez-Tenorio A, Maldonado-Rodriquez R, Doktycz MJ, Fleming JT, Beattie KL. Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization. Nucleic Acids Res 2003;31:779–789. [PubMed: 12527788]
6. Gingeras TR, Ghandour G, Wang E, Berno A, Small PM, Drobniewski F, Alland D, Desmond E, Holodniy M, Drenkow J. Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. Genome Res 1998;8:435–448. [PubMed: 9582189]

7. Kingsley MT, Straub TM, Call DR, Daly DS, Wunschel SC, Chandler DP. Fingerprinting closely related *Xanthomonas* pathovars with random nonamer oligonucleotide microarrays. Appl Environ Microbiol 2002;68:6361–6370. [PubMed: 12450861]

8. Willse A, Straub TM, Wunschel SC, Small JA, Call DR, Daly DS, Chandler DP. Quantitiative oligonucleotide microarray fingerprinting of *Salmonella enterica* isolates. Nucleic Acids Res 2004;32:1848–1856. [PubMed: 15037662]

9. Fofanov Y, Luo Y, Katili C, Wang J, Belosludtsev Y, Powdrill T, Belapurkar C, Fofanov V, Li T-B, Chumakov S, Pettitt BM. How independent are the appearances of *n*-mers in different genomes? Bioinformatics 2004;12:2421–2428. [PubMed: 15087315]

10. Vainrub, A.; Li, T-B.; Fofanov, Y.; Pettitt, BM. Biomedical Technology and Device Handbook. Amsterdam: CRC Press; 2003. p. 131-166.

11. Fofanov, V.; Fofanov, Y.; Pettitt, BM. Counting array algorithms for the problem of finding appearances of all possible patterns of size *n* in a sequence. The 2002 Bioinformatics Symposium, Keck/GCC Bioinformatics Consortium; 2002. p. 14

12. Fofanov, V.; Fofanov, Y.; Pettitt, BM. Fast subsequence search using incomplete search trees. The Seventh Structural Biology Symposium, Sealy Center for Structural Biology; 2002. p. 51

13. Fofanov V, Putonti C, Chumakov S, Pettitt BM, Fofanov Y. Fast algorithm for the analysis of the presence of short oligonucleotide subsequences in genomic sequences. 2005 University of Houston Technical Report UH-CS-05-11.

APPENDICES

A. The frequency of presence of n-mers in a random sequence

Let G be a random sequence of length M of four characters {A,C,G,T}, and S be one of the 4^n possible subsequences of length n ("n-mer"). We will enumerate them, so that $\sum_{S=1}^{4^n}$ will stand for the sum with respect to all n-mers. Let $F^{M,n}(S, k)$ be the probability that S appears k times in G (*the frequency of appearance* of S in G). To define this probability one can imagine a random statistical set of N sequences of the same length. If in this set there are N_k sequences that contain S exactly k times, then

$$F^{M,n}(S,k)=\lim_{N\to\infty}\frac{N_k}{N}.$$

Let $f^{M,n}(S)$ be the probability that S is present in G (*the frequency of presence*),

$$f^{M,n}(S)=\sum_{k=1}^{M}F^{M,n}(S,k)=1-F^{M,n}(S,0).$$

All of the related statistical information is contained in the distribution of probabilities *of coappearance* of n-mers in G, $P(n\text{-mer } S \text{ appears } k_S \text{ times}, S = 1,2,\ldots,n)$,

$$\sum_{S=1}^{n_1}k_S=M-n+1\equiv M_n,$$

where M_n is a total number of n-mers in G. We will denote this distribution by:

$$P(k_1, k_2, \ldots, k_S, \ldots, k_{n_1}) = P(\{k_S\}).$$

This distribution has a multinomial form,

$$P(\{k_S\}) = M_n! \prod_{S=1}^{4^n} \frac{{p_S}^{k_S}}{k_S!}. \tag{4}$$

Here the product is taken over all configurations, such that $\sum_{S=1}^{4^n} k_S = M_n$ and p_S is the probability to find the n-mer S in G. If $n = 1$, p_S are reduced to the "elementary probabilities", p_l to find the character l in G, $l = \{A,C,T,G\}$. Assume that p_l are given, and n-mers are composed in a random manner (i.e. the characters in S are not correlated), then

$$p_S = p_A p_T \ldots p_C, \quad S = [AT \ldots C]. \tag{5}$$

One finds immediately the frequency of appearance of S in G,

$$F^{M,n}(S,k) = \sum_{\{k_T, T \neq S\}} P(\{k_T\}) = \frac{M_n! p_S^k q_S^{M_n - k}}{k!(M_n - k)!},$$
$$q_S = 1 - p_S$$

The mean number of appearances, $\bar{k}_S$, the variance $\sigma_S^2 = \overline{k_S^2} - (\bar{k}_S)^2$, covariance $\sigma_{ST}^2 = \overline{k_S k_T} - \bar{k}_S \bar{k}_T$, and the correlation coefficient $C_{ST} = \sigma_{ST}^2 / \sigma_S \sigma_T$ are given as follows,

$$\bar{k}_S = M_n p_S,$$
$$\sigma_S^2 = M_n p_S q_S,$$
$$\sigma_{ST}^2 = -M_n p_S p_T,$$
$$C_{ST} = -\sqrt{\frac{p_S p_T}{q_S q_T}}.$$

Let us find the probability of presence, $f^{M,n}(S) = 1 - F^{M,n}(S, 0) = 1 - (1 - p_S)^{M_n}$. One may use the mean number of appearances as a new variable, $y = M_n p_S = \overline{k_S}$, and consider the common Poisson limit of the Bernoulli distribution:

$$f^{M,n}(S) = 1 - \left(1 - \frac{y}{M_n}\right)^{M_n} \to 1 - e^{-y}$$
$$y = M_n p_S. \tag{6}$$

In the case of equal probabilities $p_l = 1/4$, we have the homogeneous distribution of n-mers, $p_l = 1/4^n$ and we arrive at the frequency of presence for a random genome with equal nucleotide probabilities [9].

In the microbial genomes considered here nucleotides A and T appear with nearly equal probabilities; the same is true for nucleotides C and G, as seen in Figure 2. However, nucleotides A and C appear with different probabilities:

$$p_G = p_C = p_1$$
$$p_A = p_T = 1/2 - p_1.$$

The distribution of the values of the probability $p_G + p_C = 2p_1$ to find G and C nucleotides in different microbial genomes is shown in Figure 2b.

The probability to find the n-mer S depends on the GC content $m = n_G + n_C$

$$p_S = p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G} = p_1^m \left(\tfrac{1}{2} - p_1\right)^{n-m}. \quad (7)$$

The frequency of presence, $f(S)$, is constant for a given GC content m; however $f(S)$ depends on m and on the probability p_1. There are N_m different n-mers of GC content m,

$$N_m = \frac{2^n n!}{m!(n-m)!}, \quad \sum_{m=0}^{n} N_m = 4^m. \quad (8)$$

It means that the average number of n-mers of GC content m in a random genome G of size M is given as

$$M'_m = f(S_m, p_1) N_m, \quad (9)$$

where f and N_m are given by eqns 6, 7 and 8. When $p_1 = 1/4$, we recover the homogeneous case,

$$\sum_m M'_m = f(S) \sum_m N_m = 4^n f(S) = M',$$

where M' is a total number of different n-mers in G.

B. Intersections of two random genomes

Consider two random genomes G_1 and G_2 of sizes M_1 and M_2 and probabilities of G and C nucleotides p_1 and p_2, correspondingly. The probability that the n-mer of GC content m be present in both genomes is $f^{M_1}(S_m, p_1)\, f^{M_2}(S_m, p_2)$.

We find the total number of n-mers that are expected to be present in both genomes ("the size of the intersection of G_1 and G_2"), multiplying by N_m and summing up the terms with different values of m:

$$\sum_{m=0}^{n} N_m f^{M_1}(S_m, p_1) f^{M_2}(S_m, p_2).$$

C. Estimation of the probability of error

Let us estimate the probability to make an error in discriminating organisms by their fingerprints on a random array that consists of L n-mers. Assume that we need to discriminate between the two genomes G_1 and G_2 of corresponding sizes M_1 and M_2. Let G_1 (G_2) contain N_1 (N_2) different n-mers and N_{12} n-mers are present simultaneously in both genomes (this is the size of intersection $G_1 \cap G_2$ of the n-mer contents of G_1 and G_2, see Fig. 8). The union $G_1 \cup G_2$ contains $N_1 + N_2 - N_{12}$ n-mers. Probabilities to find an arbitrary n-mer S, correspondingly, in G_1, G_2, $G_1 \cap G_2$ and $G_1 \cup G_2$ are $p_1 = \frac{N_1}{4^n}$, $p_2 = \frac{N_2}{4^n}$, $p_{12} = \frac{N_{12}}{4^n}$, $p_1 + p_2 - p_{12}$, respectively.

Consider a joint fingerprint of $G_1 \cup G_2$. Two genomes can be distinguished if this fingerprint contains n-mers from the set $B = G_1 \cup G_2 - G_1 \cap G_2$; if all of the n-mers in the joint fingerprint belong to the sets $A = G_1 \cap G_2$ or $C = 1 - G_1 \cup G_2$, the two genomes cannot be distinguished (the set of all n-mers = $A + B + C$). The probability that an arbitrary n-mer $S \notin B$ is $1 - p_1 - p_2 + 2p_{12}$ and the probability that the random array fails to distinguish the two genomes (i.e. all of the n-mers in the joint fingerprint find themselves in A or C) is equal to $\varepsilon = (1 - p_1 - p_2 + 2\, p_{12})^L$.

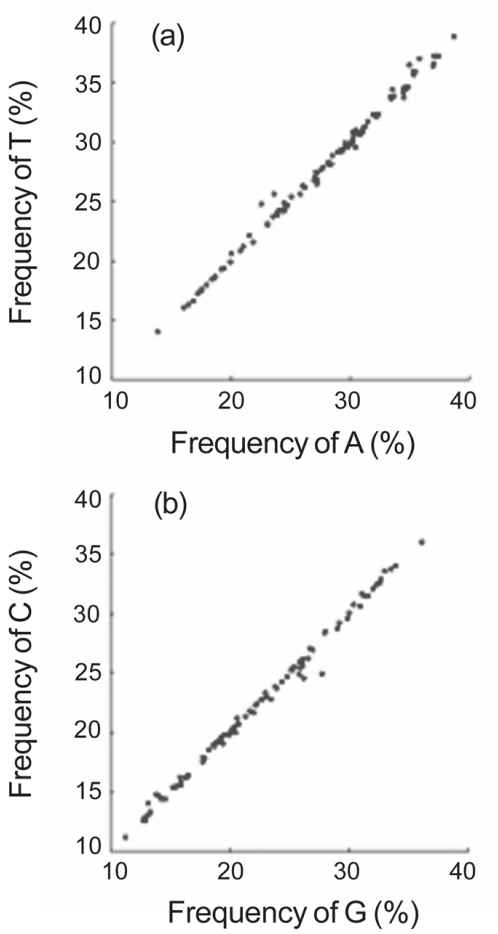

Figure 1.
Co-presence of (a) A versus T and (b) G versus C nucleotides in microbial genomes. (Only one strand of every genome is considered.)

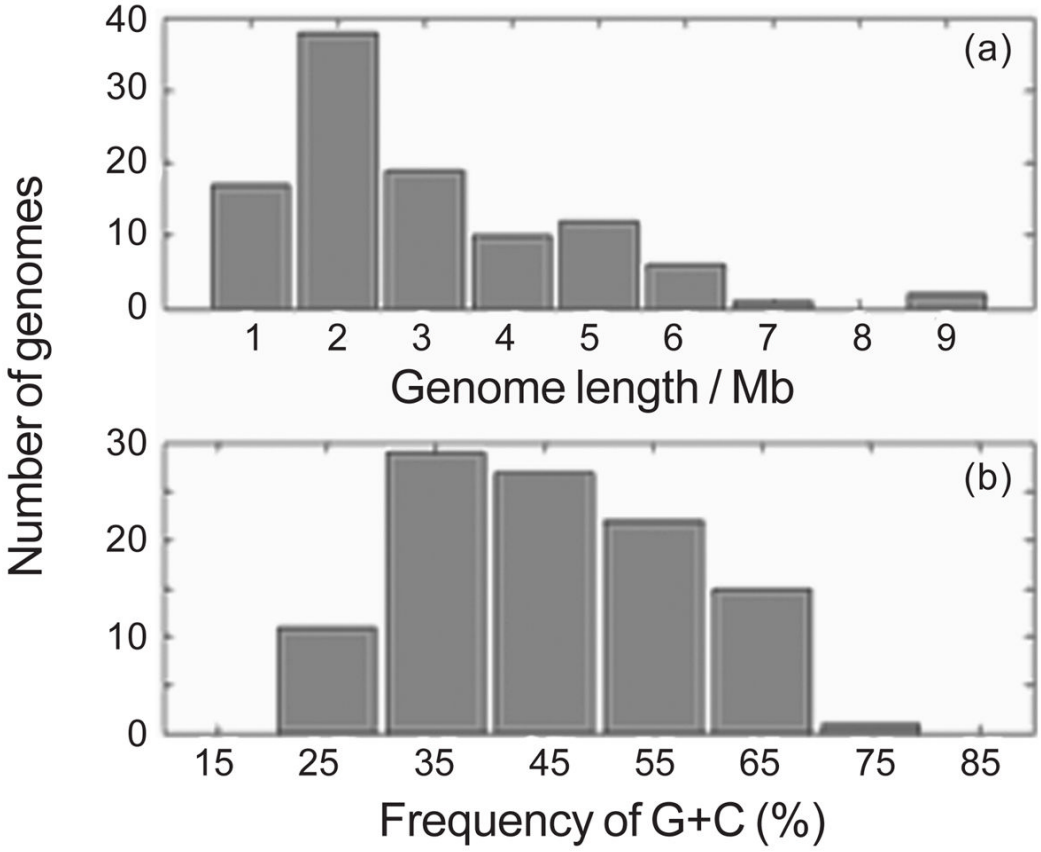

Figure 2.
Distribution of (a) genome lengths and (b) GC contents (binned into 10% ranges) for microbial genomes. The horizontal axes represent (a) the genome length and (b) the probability $p_G + p_C$ of appearance of G and C nucleotides in a genome.

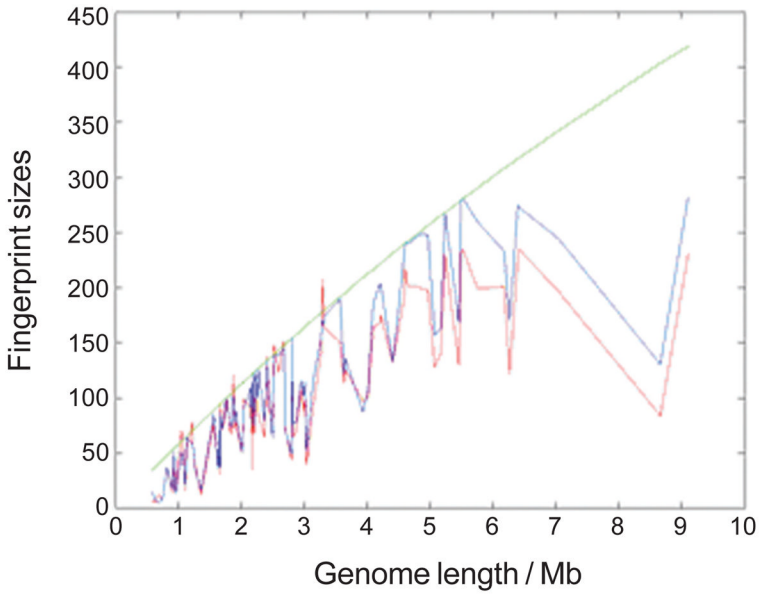

Figure 3.

Sizes of microbial fingerprints on an array of $L = 1000$ randomly chosen 12-mers with GC content $m = 6$. Every point represents a genome, the horizontal axis being the genome length. Fingerprint sizes for 105 real genomes (blue) are compared to the ones for random-sequence genomes with the same length and GC content (red) and for random genomes with the same length and uniform nucleotide distribution $p_A = p_C = p_G = p_T = 1/4$ (green).

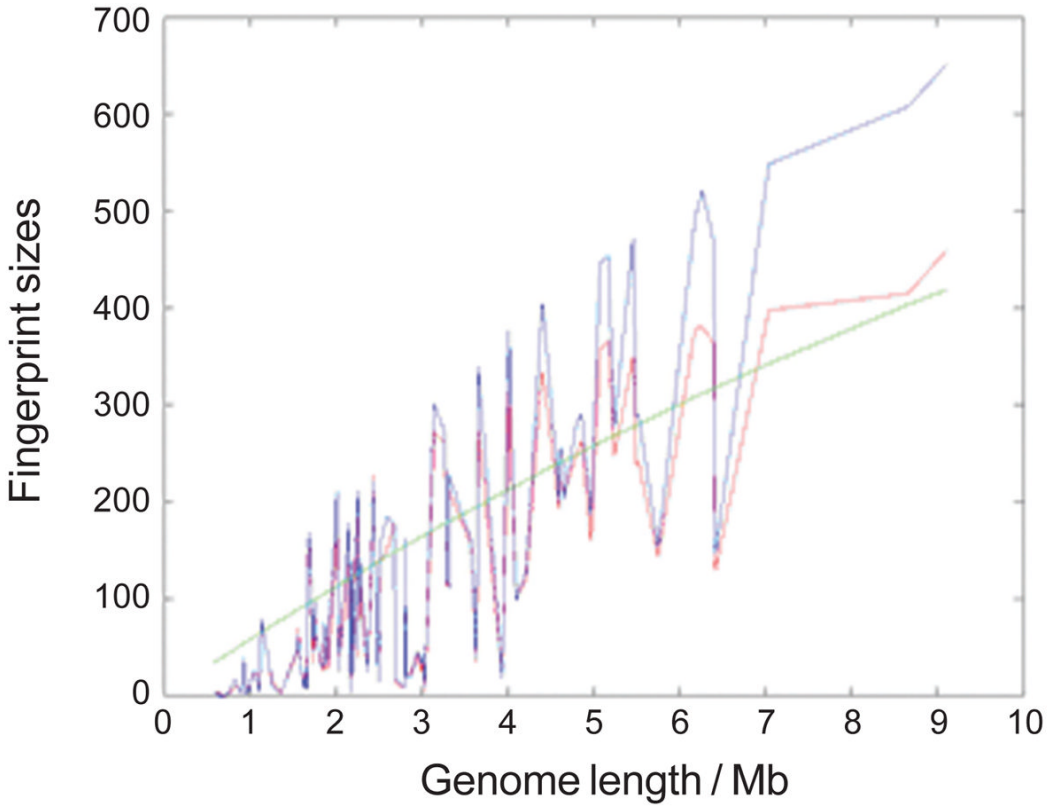

Figure 4.
The same as Fig. 3 except that the GC content $m = 8$.

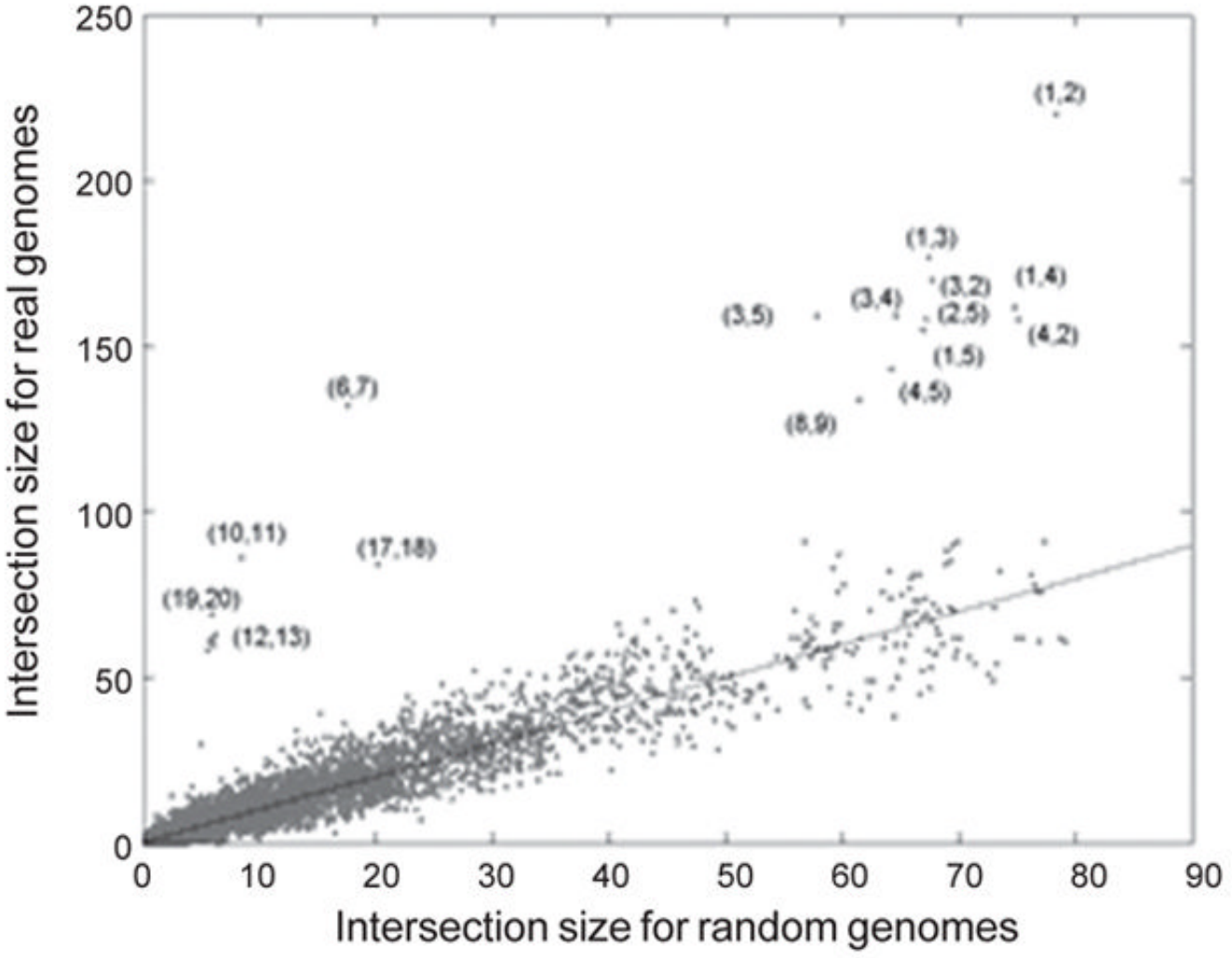

Figure 5.

Intersections of fingerprints for 105 microbial genomes on an array of $L = 1000$ randomly chosen 12-mers with GC content $m = 6$. Intersection sizes for real genomes are shown against the intersection sizes for random genomes with the same length and GC content. For most of the genome pairs intersection sizes are close to the ones for random "genomes" (i.e. the points on the scatter plot are close to the solid line ($x = y$). Deviations are observed for closely related species; some of them are marked and listed in Table 1.

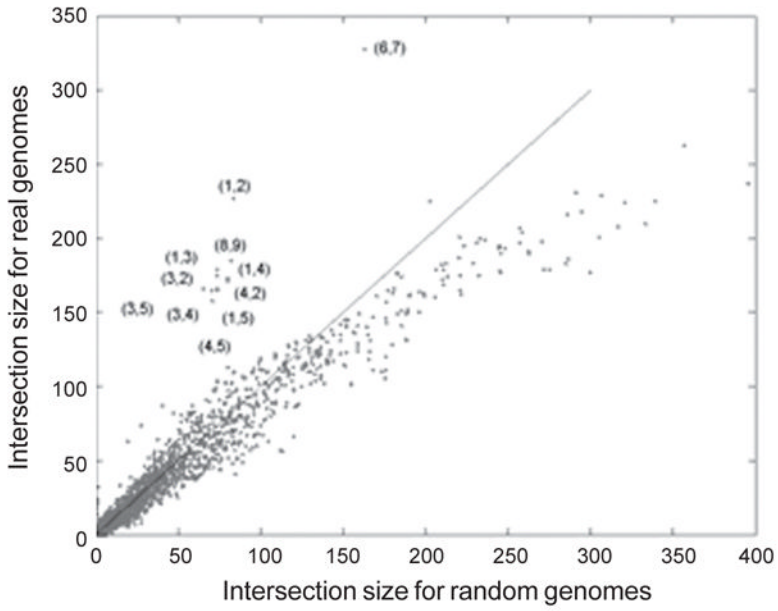

Figure 6.
The same as Fig. 5 but for GC content $m = 8$.

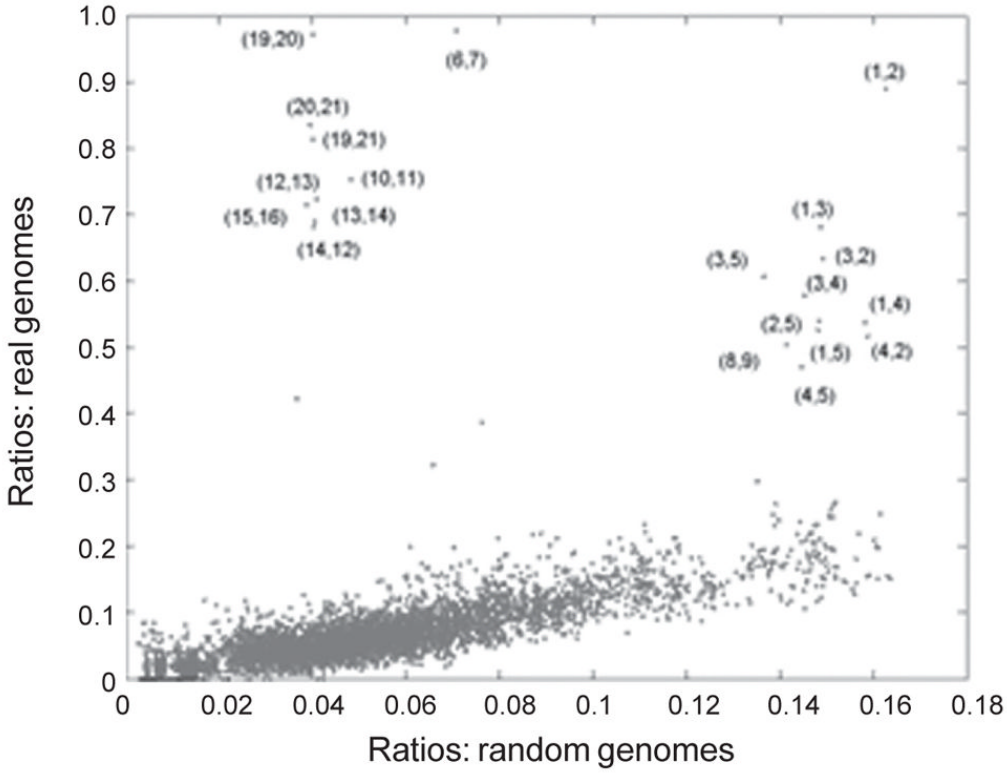

Figure 7.

Ratios, (size of intersection of positive probes)/(size of union of positive probes), for pairs of fingerprints of 105 microbial genomes on the given chip of $L = 1000$ and $m = 6$. Results for real genomes are shown against the ones for random "genomes" with the same length and GC content. Close relatives appear far from the rest of the points. Even the most similar pairs (strains of the same species; designated by their genome-list indices) can be distinguished.

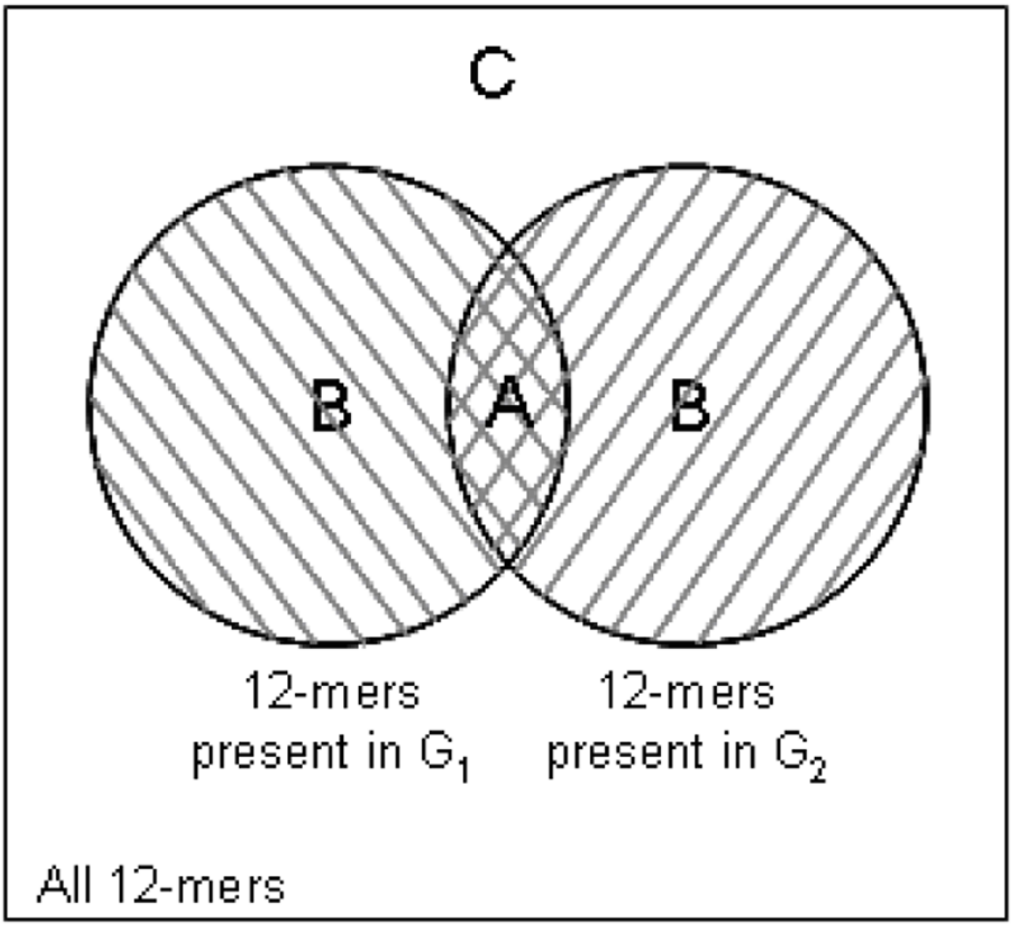

Figure 8.
The *n*-mer content of the intersection of two genomes.

Table 1

Closely related species as they appear in Figs 5–7.

1	BA000007	*Escherichia coli O157:H7*
2	AE005174	*Escherichia coli O157:H7*
3	U00096	*Escherichia coli K-12 MG1655*
4	AE014075	*Escherichia coli CFT073*
5	AE005674	*Shigella flexneri 2a str 301*
6	AL123456	*Mycobacterium tuberculosis*
7	AE000516	*Mycobacterium tuberculosis CDC1551*
8	AE006468	*Salmonella typhimurium LT2*
9	AL513382	*Salmonella typhi strain CT18*
10	AE005672	*Streptococcus pneumoniae*
11	AE007317	*Streptococcus pneumoniae R6*
12	AE009949	*Streptococcus pyogenes strain MGAS8232*
13	AE014074	*Streptococcus pyogenes MGAS315*
14	AE004092	*Streptococcus pyogenes strain SF370 serotype M1*
15	AL732656	*Streptococcus agalactiae NEM316*
16	AE009948	*Streptococcus agalactiae*
17	NC004556	*Xylella fastidiosa Temecula1*
18	AE003849	*Xylella fastidiosa*
19	BA000017	*Staphylococcus aureus strain Mu50*
20	BA000018	*Staphylococcus aureus strain N315*
21	BA000033	*Staphylococcus aureus MW*