Identification of horizontally acquired DNA using genome signature analysis

Stephen A. Jarvis[1,*], Jason S. Mirsky[1], John F. Peden[2] and Nigel J. Saunders[3†]

Programming Research Group, Oxford University Computing Laboratory[1],
University of Oxford, Oxford OX1 3QD, UK.
MRC Molecular Haematology Group[2], Molecular Infectious Diseases Group[3],
Institute of Molecular Medicine,
University of Oxford, Oxford OX3 9DS, UK.

September 1, 2000

Abstract

Motivation: Identification of regions with untypical percentage G+C composition and dinucleotide signatures are two genome analysis techniques used in the identification of horizontally acquired DNA. We describe a generic program framework for performing both types of analysis in linear time. The approach is extended for length > 2 oligonucleotide signatures. Using the derived program we test some of the conclusions of Karlin and Burge, the primary exponents of these techniques.
Results: We demonstrate that no single method of signature analysis is sufficient for the complete identification of horizontally acquired DNA. Consequently we produce a fast program - and a robust methodology - for the production and employment genome signatures in the identification of horizontally acquired DNA.
Availability: Software available from first author.
Keywords: DNA; genome; signature; algorithms.

*now at: Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK. Email: *stephen.jarvis@dcs.warwick.ac.uk*. To whom correspondence should be addressed

†current address: Sir William Dunn School of Pathology, University of Oxford, Oxford. OX1 3RE

Introduction

In addition to the vertical transmission of adaptive mutations, horizontal transfer of genes between unrelated species is increasingly recognised as an important component of evolution [Doolittle, 2000]. Most if not all species include material which has been acquired in this way. The scope for this type of genetic exchange is limited in higher organisms in which the germ cells are sequestered and are not normally exposed to foreign DNA. However, in bacteria, in which the cells that form the origins of new populations are exposed to DNA in the environment, horizontal exchange of DNA is an ongoing and important process [Arber, 1993, Lawrence *et al.*, 1998, Maddox, 1998, Woese, 1998, Brown and Doolittle, 1999, Doolittle, 1999, Forterre and Philippe, 1999, Jain *et al.*, 1999, Martin, 1999, Nelson *et al.*, 1999, Faguy and Doolittle, 2000, Moreire, 2000]. For example, it has been estimated that 18% of the genes in *E. coli* have been acquired by horizontal exchange since its divergence from *Salmonella* spp. approximately 100 million years ago [Lawrence *et al.*, 1998].

Horizontal exchange is recognised as being important in the evolution of bacterial virulence, the most widely acknowledged examples of which are so-called 'pathogenicity islands'. The typical characteristics of these regions of DNA are: a %G+C composition which differs from that of the rest of the genome; terminal inverted repeats; proximity to tRNA genes [Hou, 1999] and association with transposases which may be intact or present as remnants. These islands often contain groups of genes with functions which can readily be invoked in making a species more virulent, such as toxin production and secretion [Groisman and Ochman, 1996, Hacker *et al.*, 1997, Alfano *et al.*, 2000]. Usually the origin of the transferred DNA is not known and the source of these genes - so significant in virulence - remains an interesting enigma. In a small number of instances, however, the likely origin of horizontally transferred genes can be inferred with some certainty [Vasquez *et al.*, 1995, Kroll *et al.*, 1998, Saunders *et al.*, 1999].

In order to become established within a bacterial population, an acquired gene (or other gene with which it is associated) must lead to an increase in bacterial fitness. This takes place through one or more of the following factors: increasing transmissibility, increasing the size of the colonising population, establishing new niches in which the organism can grow, or evading host immune responses. This increase in fitness must then be sufficient for the sequence to become prevalent in the population to which it has been transferred. This occurs either through clonal replacement of the previous population that did not have the gene or by further horizontal transfer of the gene within the new species to which it has been transferred. The processes which might be expected to be involved in increased fitness are broadly similar to those involved in virulence. It is therefore important to be able to identify sequences which have been acquired by horizontal transfer in order to understand the evolutionary history of a species and, in the context of bacterial pathogens, to identify genes potentially associated with virulence.

DNA from different genetic backgrounds (e.g., bacterial species) has different characteristics. These are a product of the codons which are preferentially used by the particular species and the environmental conditions in which it grows, as well as the types of error to which the DNA replication enzymes are prone and differences in the ability of each enzyme to correct mistakes. This makes it possible to identify sections of DNA which are derived from different genetic backgrounds by means of their untypical base composition. DNA which is incorporated following horizontal transfer will be exposed to the same pressures which gave the recipient its characteristic composition. As a result, over time the acquired DNA will begin to conform to the average DNA composition of the recipient by a process called amelioration [Lawrence and Ochman, 1997]. However, the rate of amelioration is sufficiently slow that it is possible to identify regions of DNA which have been horizontally acquired. The extent to which this sequence is untypical therefore depends on both the extent of the difference between the donor and recipient species and the period of time which has elapsed since acquisition.

The first method for the identification of regions of potentially horizontally transferred DNA is based on %G+C characteristics of the sequence. This is useful because the %G+C within prokaryotic genomes is relatively homogeneous within a species but differs between unrelated species [Karlin *et al.*, 1998]. Significantly different %G+C content in sections of DNA therefore indicates possible horizontal acquisition [Karlin *et al.*, 1998].

Experimental studies showed that the dinucleotide odds ratios were similar in most organisms for the bulk of their DNA and essentially the same for different parts of the same genome [Josse *et al.*, 1961]. Looking at sequences derived from different sequence backgrounds revealed that more closely related species tended to have more similar dinucleotide compositions than unrelated species [Karlin *et al.*, 1994.a, Karlin *et al.*, 1994.b, Karlin *et al.*, 1994.c].

Karlin and his colleagues noted that the species differences in base usage were more complex and that, in addition to differences in %G+C composition, different species used characteristic patterns of dinucleotide pairs within their genomes. This observation provided the basis for the second approach in which the DNA is seen not as a string of single bases but as a sequence of pairs of dinucleotides: the sequence ACGTG, for example, would be considered as a sequence of dinucleotide pairs AC, CG, GT and TG. The percentage of each of the 16 dinucleotide pairs was found to differ between species and although the reasons for these differences are not fully understood, this provides a second method for identifying horizontally acquired DNA. The method is known as dinucleotide signature (DNS) analysis [Karlin and Burge, 1995].

In this paper:

- Karlin's approach to %G+C content and DNS analysis is examined and extensions to his methods are proposed.

- A generic program for the linear computation of n-oligonucleotide signatures is developed.

- Karlin and Burge's observation that n-oligonucleotide signatures are highly implied by the DNS is confirmed. However, it is also demonstrated that if different window sizes and word lengths are used, n-oligonucleotide signatures point to different, potentially relevant, areas from the DNS graphs.

- A comparison is made of the different methods for identifying horizontally acquired DNA; conclusions are drawn and supporting results, which describe how the methods can be used, are presented.

Systems and methods

Computational analysis

Asymptotic analysis is a way of comparing the performance of different computational algorithms. This comparison is used to select one computational approach, which may provide a more efficient solution, over another.

The big-O notation describes the upper bound of the asymptotic growth of a function. For instance, if an algorithm takes $\frac{1}{3}n^2 - 2n + 4$ steps, it is of $O(n^2)$, since it is bounded from above by a function which takes n^2 steps times some constant [Kaldewaij, 1990].

Karlin's methods for finding horizontally acquired DNA rely on comparing DNA with itself by means of mathematical calculation. The chromosomes of bacteria are typically composed of 0.5 to 5 million base pairs (Mb). Performing calculations on the DNA sequence (of length n) requires a large amount of computation: first the complete sequence statistics are obtained, $O(n)$; then individual segments (windows) are statistically analysed and compared with the complete sequence, $O(n^2)$. A straightforward refinement would therefore produce an implementation which was bounded by $O(n^2)$ steps.

Number of steps (bounded by O)	Estimated execution time
$O(\log 2\ n)$	0.1 *seconds*
$O(\sqrt{n})$	5 *seconds*
$O(n)$	1 *hour* 15 *minutes*
$O(n^2)$	450 *years*

Figure 1: Length of time n computations take when $n = 5,000,000$ and a single computation takes 1 millisecond. Table adapted from [Kaldewaij, 1990].

Figure 1 shows the relationship between asymptotic analysis and computation time.

Optimisations to programs can of course be made and computers work at a far greater rate than the example suggests. However, the table demonstrates that a difference in the order of complexity of an algorithm is directly proportional to the runtime. Reducing an algorithm from an $O(n^2)$ solution to a $O(n)$ solution will make a considerable difference to the number (and variety) of results which can feasibly be obtained.

Methods of genome analysis

%G+C content

Calculating the %G+C content of an organism's DNA is a simple method for finding regions of genetic interest. %G+C content scores are calculated for segments (windows) of DNA by adding the frequency (number divided by window length) of G to the frequency of C. A plot of %G+C content is derived by sliding the window over the complete genome sequence and plotting the results.

Dinucleotide signatures

Dinucleotide signatures can be used to show areas where pairs of bases are clustered more frequently than if they were distributed across the sequence by chance. A DNS graph is based on a calculation of odds ratios for each of the 16 dinucleotides. This odds ratio (probability) is calculated by taking the frequency of a dinucleotide and dividing it by the expected probability of finding the dinucleotide within a particular sequence:

$$p_{xy} = \frac{f_{xy}}{f_x * f_y} \tag{1}$$

f_x is the frequency of nucleotide x within the sequence and f_{xy} is the frequency of the dinucleotide xy within the (circular) sequence of DNA:

$$f_{xy} = \frac{\#xy}{\#dinucleotides} \tag{2}$$

where $\#dinucleotides = SequenceLength - 1$

p_{xy} represents the odds ratio for dinucleotides in single-stranded DNA. To calculate the odds ratio for double-stranded DNA, written $\overset{*}{p}_{xy}$, the sequence and its reverse complement are concatenated [Burge *et al.*, 1992]. The calculation remains the same, but with a sequence twice the length of the original.

To find areas of possible horizontally acquired DNA, a window (f) (of some chosen size – 50 Kb for example [Karlin and Burge, 1995]) is created. The dinucleotide probabilities of this window are then compared with the dinucleotide probabilities of the overall sequence (g). This is done for each of the 16 dinucleotides and the normalised results are added. The result is an overall dinucleotide difference between the window f and the sequence g.

$$\overset{*}{\delta}(f,g) = (\frac{1}{16}) \sum_{xy}^{16} \mid \overset{*}{p}_{xy}(f) - \overset{*}{p}_{xy}(g) \mid \tag{3}$$

Using the same mechanics as those used in %G+C content signatures, Karlin and Burge suggest repeatedly sliding the window one position along the sequence and repeating the difference calculation until finally the sequence is completely covered by windows. Graphing the results shows the regions of largest difference, which are candidates for horizontal acquisition.

Oligonucleotide signatures

Just as greater information and resolution is obtained when dinucleotide sequence composition is considered rather than single base composition, analyzing sequences on the basis of even longer motifs may provide additional useful information. Trinucleotide composition addresses codon usage in expressed portions of open reading frames. However, it may be that preferences for particular pairs or more of codons, mutational processes which favour or select against longer sequences, or processes which affect the composition of the non-transcribed strand will influence the sequence composition in additional ways. The method has therefore been extended so that signatures can be determined for longer motifs [Karlin *et al.*, 1997, Mirsky, 1999].

Dinucleotide analysis is easily extendible to oligonucleotide signatures of length n.

$$\overset{*}{\delta}(f,g) = (\frac{1}{4^l}) \sum_{i:s}^{4^l} \mid \overset{*}{p}_i(f) - \overset{*}{p}_i(g) \mid \tag{4}$$

where l is the length of oligonucleotides which comprise the signature, s is the set of all permutations of length l and i is one such permutation.

In addition, the variance v and standard deviation sd can be determined for the absolute differences of the p values at each point and for the δ values across the entire sequence.

$$v = \frac{n * \sum x^2 - (\sum x)^2}{n^2} \tag{5}$$

To calculate the variance for $\overset{*}{\delta}$, each $\overset{*}{\delta}$ value corresponds to a x and the number of $\overset{*}{\delta}$ values (the sequence length) corresponds to n. To calculate the variance of the p values, $x = p$ and $n = 4^l$. The standard deviation is simply the sum of the variance values, $sd = \sqrt{v}$.

When the entire population is not known, that is if the window slides by more than one base between calculations, an estimate of the variance is calculated:

$$v = \frac{n * \sum x^2 - (\sum x)^2}{n(n-1)} \tag{6}$$

Implementation

The same procedure is used to implement %G+C content, dinucleotide and n-oligonucleotide signatures. The sliding window implementation (or similar) is mechanically equivalent for each; the primary difference between each signature is simply the formula used to calculate the points.

A generic implementation is explored.

%G+C content requires the smallest specification and is therefore most appropriate to illustrate the derivation of the program. An $O(n^2)$ solution is derived and, by a method known as loop flattening, this is converted into an $O(n)$ solution.

The $O(n)$ implementation provides the basis for a comparative study of these methods for the identification of horizontally transferred DNA.

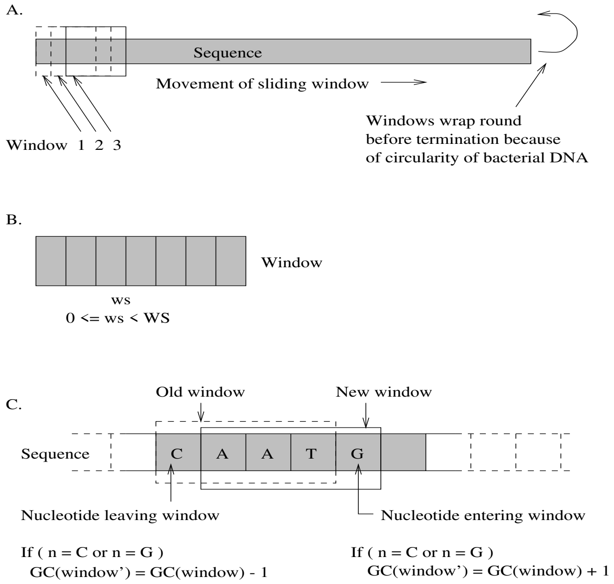

Figure 2: Derivation of a generic DNA signature algorithm: A. Generate a window for every position in the sequence by sliding the window through the sequence, the *outer loop*; B. Calculate signatures for each window, the *inner loop*; C. Flatten the inner loop for $O(N)$ solution.

Generic genome signatures

The specification of a program for %G+C content analysis (GC_{prog}) is presented in the guarded command language [Kaldewaij, 1990].

$P = \{N \geq 1\}$
[con N: int $\{N \geq 1\}$; A: array[0..N) of DNA;
  con WS: int $\{1 \leq WS \leq N\}$;
  var n: int;
  GC_{prog}
].
$Q = \{r = \forall p : 0 \leq p < N : \forall q : 0 \leq q < WS :$
  $gc = (\#i : p \leq i \leq (p+q) :$
    $A.(i \; mod \; N) = C$
    $\vee A.(i \; mod \; N) = G)/WS)\}$

The specification contains three parts: the precondition P; the part in the frame, denoted by square brackets; and the post condition Q, found after the frame.

The precondition states that the sequence must be of at least length 1 (for %G+C content this is appropriate; for larger oligonucleotides this minimum must be the length of the oligonucleotide).

The frame contains the information relevant to the development of the program: the constant N is the length of the DNA sequence A; WS is the window size; n is a variable used for traversing the sequence and GC_{prog} is simply a marker for the program itself.

The postcondition describes the result r, a list of %G+C content differences, one for each of the windows in the sequence. The modulo (*mod*) is contained within the specification to account for the circularity of the bacterial DNA. It is also noted that the results list r is written to a file as the program proceeds - this saves on RAM.

The program derivation contains three stages:

Derivation: stage 1

To calculate a window from every position in the sequence, the window slides through the sequence (cf. Karlin). This outer loop is determined by replacing the constant N by the variable n and using the post-condition $Q\{n = N\}$. This suggests starting n at 0 and setting the loop guard to $n \neq N$. This offers a solution to the outer loop of $O(N)$. See Figure 2-A.

Derivation: stage 2

The inner loop is now calculated. The inner loop corresponds to the calculation for one window. Since the inner loop will be placed within the outer loop, p can be replaced by n in the inner loop specification.

```
[  var ws, gc : int;
   ws, gc = 0, 0;
   W_prog
].
```

$$Q_i = \{t = \forall q : 0 \leq q < ws : \\ gc = ((\#i : n \leq i < (n+q) : \\ A.(i \; mod \; N) = C \\ \vee A.(i \; mod \; N) = G)/WS)\}$$

The inner loop is derived by replacing constant WS with variable ws and splitting the invariant into two chunks, one for calculating G content and the other for calculating C content. The inner block is also $O(N)$. See Figure 2-B.

Since the windows range from size 1 to $N-1$, a mathematical average for the window size is $\approx N/2$. This gives $O(WS)[WS\backslash(N/2)]$, i.e., $O(N)$ for the inner loop. As this occurs $O(N)$ times, once for each iteration of the outer loop, the complexity of the algorithm is $O(N^2)$.

Derivation: stage 3

It is possible to flatten the inner loop from an average size of $N/2$ to a constant of 2. Instead of simply counting and discarding the number of Gs and Cs in each window, it is possible to use the information from the previous window to calculate the next. This is possible by subtracting the items which leave the window and adding the items which enter the window each time the window slides. See Figure 2-C.

To calculate the new inner loop we now split the loop into two parts: the first calculates the content for the first window, $O(N)$ once only; the second calculates the remaining $N-1$ windows, constant $2 * O(N) = O(N)$. As we are adding terms rather than multiplying them, the new algorithm is $O(N)$.

Computing other signature types

The derived algorithm is generic in the sense that it provides the necessary $O(N)$ framework for whichever type of signature one is trying to program. The inner loop calculation (for %G+C content) can simply be replaced by the formula for n-oligonucleotide signatures [Mirsky, 1999].

One interesting observation should be noted for n-oligonucleotide signatures.

When nucleotides enter and leave a window, some adjustment needs to be made to the tallies for the oligonucleotides in the new window. In Figure 2 part C, for example, a count of the dinucleotides in the window requires 1 to be subtracted from the CA count and 1 to be added to the TG count.

This is an easy calculation to make, yet one must be careful about storing and accessing a list of oligonucleotide occurrences. If a linear search is used then the complexity of the overall algorithm may change; if $4^l > N$, the generic signature program will not maintain $O(N)$.

To overcome this problem a hash function is used to calculate the array position of any permutation. DNA is coded as an enumerated type, where A=0, T=1, C=2, G=3. The hash function is calculated as follows: if the oligonucleotide being searched for in the hash table is ATCA, the hash table offset is as follows:

$$0 * 4^3 + 1 * 4^2 + 2 * 4^1 + 0 * 4^0 = 0 + 16 + 8 + 0 = 24$$

The occurrence count for this oligonucleotide is therefore found at position 24 in the hash table.

The calculation is always unique, so searching the hash table is never necessary. It is also constant, so the lookup is always very quick.

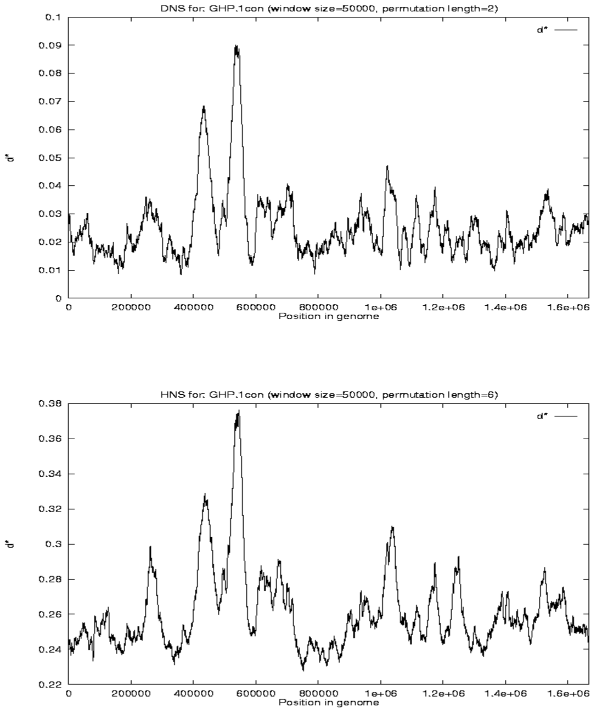

Figure 3: N-oligonucleotide signatures for *H. pylori*. Using large windows the HNS (part *b*, bottom) is highly implied from the DNS (part *a*, top).

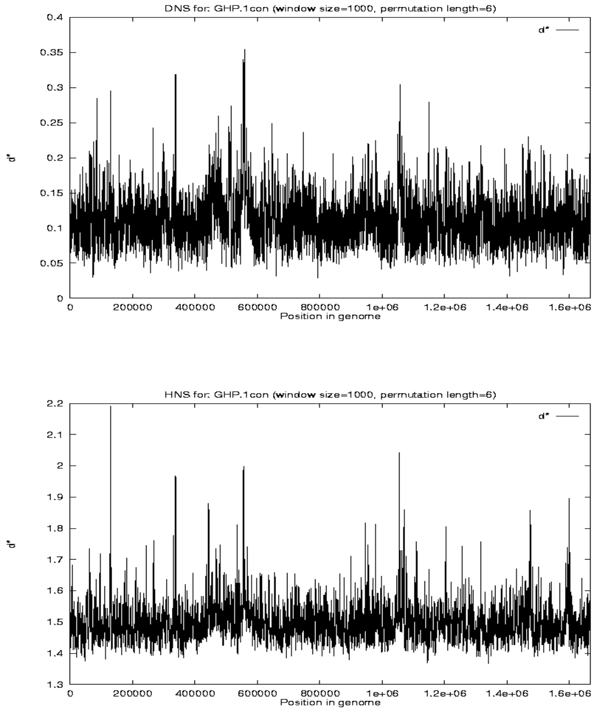

Figure 4. Signature analyses of *H. pylori*, window size 1 Kb. The first graph shows the DNS; the second shows the HNS.

This method also has the advantage of not requiring any of the oligonucleotide permutations to be stored textually, and thus wasting memory. Each oligonucleotide permutation is simply a formula based on the nucleotide content and the enumerated type values.

One last optimisation can be made. This implementation is for δ and not $\overset{*}{\delta}$, as the calculation is for single-stranded DNA. It is possible to write a function which derives the complement of the sequence and concatenates it onto the end, as suggested by Karlin and Burge (1995) and originally implemented by Burge *et al.* (1992). This turns N into $2 * N$. We prove that δ actually equals $\overset{*}{\delta}$, even though $p \neq \overset{*}{p}$. The proof of this equivalence is illustrated by Mirsky (1999). The 50% gain reduces the computation time of a derived program still further.

Discussion

Comparative results

Karlin and Burge (1995) present a DNS graph of *H. pylori*; this graph is reproduced by our derived program and is shown in Figure 3-a. Karlin and Burge document that signatures created from longer permutations are "highly implied" by the DNS graph. The length 6-oligonucleotide signature graph (HNS) for window size 50,000 (Figure 3-b) is indeed implied by the DNS results, although the results are not the same. The observation of implication does not hold uniformly for other window sizes. Different areas of interest are highlighted as the window-length and permutation-length variables are modified (Figure 4).

The results also show that $\overset{*}{\delta}$ acts as an averaging function, removing a large amount of potentially valuable data. By studying $P_{xy} - p_{xy}$ (dinucleotide probability for the whole sequence minus the probability for an individual window) for all xy permutations of length 2, regions of importance appear which are averaged out from the $\overset{*}{\delta}$ graphs.

The graphs for different permutation lengths can be compared. This is because signatures are based on odds ratios and so the number of permutations is absorbed into the ratio yielding absolute results. Significant results are identified by selection based on standard deviation thresholds from the mean. The most appropriate thresholds should be determined on an organism-by-organism basis and on the nature of the study being performed, and remain dependent on window size and permutation length.

DNS analysis using a 50,000 base window has previously been used to identify regions that have been horizontally acquired in *N. meningitidis*, using a threshold for detection of 3 standard deviations [Tettelin *et al.*, 2000]. Analysis of the *H. pylori* sequence, using the same parameters the DNS analysis, identified two regions including reading frames HP0412 to HP0465 and HP0497 to HP0573. The second region contains the cag pathogenicity island previously identified using this method [Karlin and Burge, 1995].

Using the same threshold of 3 standard deviations the HNS identified only a single region (from HP0499 to HP0578) containing the second region identified by the DNS. Using a reduced threshold of greater than 2 standard deviations the HNS also identified a region including the first identified by the DNS (from HP0409 to HP0480), and an additional region including reading frames HP0974 to HP1030.

One difficulty with this approach is that the large size of the windows means that this method only gives a general indication of where the atypical sequence is located. In the case of the cag pathogenicity island the region which had been horizontally acquired was defined by a judgement based upon interpretation of the coding regions [Karlin and Burge, 1995]. In the case of *N. meningitidis*, in which a similar analysis was performed, the limits of the regions that had been horizontally acquired were determined by comparison with an unrelated second sequenced strain [Tettelin *et al.*, 2000].

Using a window length of 1000, DNS analysis identified 177 windows of > 3 standard deviations above the mean. These correspond to 25 regions with contiguous ORFs and 2 regions which did not contain annotated coding regions. The HNS analysis identified a broadly similar number of regions, with 222 windows > 3 standard deviations above the mean. These correspond to 32 regions with contiguous ORFs (and none not containing any annotated coding regions). However, although the analyses based upon different length motifs identified 20 common reading frames (40% of DNS and 25.3% of HNS finds) the majority of the reading frames identified were unique to one or other analysis. The unique genes identified in both analyses included those which are known to be the subject to horizontal transfer such as restriction modification system genes (2 identified by DNS and 7 by HNS). It should also be noted that genes which have atypical sequence composition for reasons other than horizontal acquisition are also identified by this method, including in this analysis the 'poly E rich protein' (HP0322) and the 'histidine and glutamine-rich protein' (HP1432).

Reducing the window size increases the granularity of the results. This is associated with an increase in the specificity of this analysis, although some genes adjacent to those that generate divergent results will still be included due to the sliding window nature of this method. The results concerning the cag pathogenicity island proved interesting. Both 1000 bp window size analyses identified HP0527 and HP0528 (coding for cag pathogenicity island proteins 7 and 8). However, the DNS also identified HP0522, HP0523, HP0531, HP0532, HP0533 and HP0534 (encoding cag pathogenicity island proteins 3, 4, 11, 12, 13 and a hypothetical protein, respectively); while the HNS identified HP0527 and HP0528 (encoding cag pathogenicity island proteins 15 and 16) as well as two other reading frames within the region of divergence but not part of the recognized pathogenicity island. These results indicate that although the overall patterns of the longer window analyses are similar this can be the consequence of the presence of different signals within the scanned areas and that the specificity and sensitivity of the different word length analyses differ.

Karlin and Burge state that the area of most significance in the 50 Kb window dinucleotide signature points to a pathogenicity island [Karlin and Burge, 1995]. This is supported by the 50 Kb-window DNS and %G+C content results. However, the analysis using shorter windows indicates that some of the genes within this region, even within the cag pathogenicity island itself, are not divergent by 1 standard deviation (see Figure 4). On the basis of the signature analysis it cannot be concluded that all of the genes within this region have features that suggest that they have been horizontally acquired. While a 1 Kb window implies the results of the 50 Kb window, since the larger window merely averages the results of the smaller windows it comprises, the reverse is not true. Examination of the region identified by the 50 Kb DNS using the smaller window analysis allows us to conclude that the large peak surrounding the pathogenicity island is in part identifying regions of atypical associated DNA, and that the abnormal regions do not comprise all the recognised genes that compose the cag pathogenicity island. Signature analysis with a large window (e.g. 50 Kb) cannot be relied upon for finding short regions of atypical sequence (e.g. 2 Kb) in a genome.

Recently, Lawrence and Ochman mined the DNA of *E. coli* for horizontally acquired regions [Lawrence *et al.*, 1998]. Using a codon bias technique, they determined that 18% of the genes in *E. coli* (approximately 10% of the sequence) have been horizontally acquired since its divergence from *Salmonellae*. A similar proportion of untypical sequence in *E. coli* is identified using signature analysis. A length-2 oligonucleotide signature and a length-6 oligonucleotide signature showed that 14% and 12% respectively of the windows in each were at least 1 standard deviation above the mean.

Conclusions

This paper describes new software for the identification of horizontally acquired DNA. Programs have been developed which perform dinucleotide and n-oligonucleotide signatures and which include methods to plot individual probabilities and to calculate standard deviations and means for use in confidence estimates. The development of a generic $O(n)$ algorithm for these tasks is documented. This has considerable benefit over an $O(n^2)$ solution. The speed at which the resulting program executes means that methods of signature analysis can be compared using large datasets and previous conclusions can be tested.

Testing these signature analysis methods on genome sequence data has revealed that %G+C content alone is not a reliable method for identifying horizontally acquired DNA. The variance that implies horizontal transfer is not always reflected in %G+C content and therefore we recommend the use of %G+C content only to support data from one of the other documented methods.

DNS analysis does identify sequences consistent with those thought to be horizontally transferred. The generation of length-n>2 oligonucleotide signatures also provides useful results. Karlin's results suggesting that the n-oligonucleotide signatures are highly implied by the dinucleotide signature are confirmed for a 50 Kb window. However, with different window sizes, the n-oligonucleotide signatures often point to other, potentially relevant, areas from the dinucleotide signature graphs. In addition, we find that plotting individual differences in probabilities, before averaging, results in a finer granularity of results. Often an average cannot distinguish between a high variation in an individual permutation which is offset by low variations in the other permutations, on the one hand, and moderate difference in all permutations on the other.

Our comparison of methods leads us to conclude that no single method of signature analysis is sufficient for the complete identification of horizontally acquired DNA. There is some overlap between the results, but different approaches contribute valuable additional results that any one method would miss.

Acknowledgements

Nigel Saunders is supported by a Welcome Trust fellowship in medical microbiology.

References

Alfano,J.R., Charkowski,A.O., Deng,W.L., Badel, J.L., Petnicki-Ocwieja,T., van Dijk,K. and Collmer, A. (2000) The *Pseudomonas syringae* hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proc. Natl. Acad. Sci. USA.* **97** 4856-61.

Arber,W. (1993) Evolution of prokaryotic genomes. *Gene* **135** 49-56.

Brown,J.R. and Doolittle,W.F. (1999) Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. *J. Mol. Evol.* **49** 485-95.

Burge,C., Campbell,A.M., Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* **89** 1358-1362.

Doolittle,W,F. (1999) Lateral genomics. *Trends Cell Biol.* **9** M5-8.

Doolittle,W.F. (2000) Uprooting the tree of life. *Sci. Am.* **282** 90-5.

Faguy,D.M. and Doolittle,W.F. (2000) Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. *Trends Genet.* **16** 196-7.

Forterre,P. and Philippe,H. (1999) Where is the root of the universal tree of life? *Bioessays* **21** 871-9.

Groisman,E.A. and Ochman,H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87** 791-4.

Hacker.J., Blum-Oehler,G., Muhldorfer,I. and Tschape,H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* **23** 1089-1097.

Hou,Y.M. (1999) Transfer RNAs and pathogenicity islands. *Trends Biochem. Sci.* **24** 295-8.

Jain et al. (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. USA.* **96** 3801-?.

Josse,K., Kaiser,A.D. and Kornberg,A. (1961) Enzymatic synthesis of deoxyribonucleic acid VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.* **263** 864-875.

Kaldewaij,A. (1990) *Programming: The Derivation of Algorithms.* Prentice Hall Int. Series in Comp. Science, Prentice Hall, New York.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Tren. Genet.* **11** 283-290.

Karlin,S., Campbell,A.M. and Mrazek,J. (1998) Comparative DNA analysis across diverse genomes. *Annual Rev. Genet.* **32** 185-225.

Karlin,S., Ladunga,I., Blaisdell,B.E. (1994) Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. USA* **91** 12837-12841.

Karlin,S. and Ladunga,I. (1994) Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* **20** 12832-12836.

Karlin,S., Mocarski,E.S. and Schachtel,G.A. (1994) Molecular evolution of herpesviruses: genomic and protein sequence comparisons. *J. Virol.* **68** 1886-1902.

Karlin,S., Mrazek,J. and Campbell,A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J. Bact.* **179** 3899-3913.

Kroll,J.S., Wilks,K.E., Farrant,J.L. and Langford,P.R. (1998) Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. *Proc. Natl. Acad. Sci. USA* **95** 12381-5.

Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44** 383-397.

Lawrence,J.G., Ochman,H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* **95** 9413-7.

Maddox,J. (1998) *What Remains to Be Discovered.*

TheFree Press, Simon & Schuster, New York.

Martin,W. (1999) Mosaic bacterial chromosomes: a challenge enroute to a tree of genomes. *BioEssays* **21** 99-104.

Mirsky,J. (1999) *Genome Analysis Methodologies for the Identification of Horizontally Acquired DNA.* Oxford University Computing Laboratory M.Sc. thesis.

Moriere (2000) Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Mol. Microbiol.* **35** 1-5.

Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A., McDonald,L., Utterback, T.R., Malek,J.A., Linher,K.D., Garrett,M.M., Stewart,A.M., Cotton,M.D., Pratt,M.S., Phillips,C.A., Richardson,D., Heidelberg,J., Sutton, G.G., Fleischmann,R.D., Eisen,J.A., Fraser,C.M., et al. (1999) Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima.* *Nature* **399** 323-9.

Saunders,N.J., Hood,D.W. and Moxon,E.R. (1999) Bacterial evolution: bacteria play pass the gene. *Curr. Biol* **9** R180-3.

Tettelin,H., Saunders,N.J., Heidelberg,J., Jeffries,A.C., Nelson,K.E., Eisen,J.A., Ketchum,K.A., Hood,D.W., Peden,J.F., Dodson,R.J., Nelson,W.C., Gwinn,M.L., DeBoy,R., Peterson,J.D., Hickey,E.K., Haft,D.H., Salzberg,S.L., White,O., Fleischmann,R.D., Dougherty,B.A., Mason,T., Ciecko,A., Parksey,D.S., Blair,E., Cittone,H., Clark,E.B., Cotton,M.D., Utterback,T.R., Khouri,H., Qin,H., Vamathevan,J., Gill,J., Scarlato,V., Masignani,V., Pizza,M., Grandi,G., Sun,L., Smith,H.O., Fraser,C.M., Moxon,E.R., Rappuoli,R. and Venter,J.C. (2000) The complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **278** 1809-1815.

Vasquez,J.A., Berron,S., O'Rourke,M., Carpenter,G., Feil,E., Smith,N.H. and Spratt,B.G. (1995) Interspecies recombination in nature: a meningococcus that has acquired a gonococcal PIB porin. *Mol. Microbiol.* **15** 1001-1007.

Woese,C. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. USA* **95** 6854-9.

References

[Karlin and Burge, 1995] Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Tren. Genet.*, **11** 283-290.

[Doolittle, 2000] Doolittle,W.F. (2000) Uprooting the tree of life. *Sci. Am.* **282** 90-5.

[Arber, 1993] Arber,W. (1993) Evolution of prokaryotic genomes. *Gene* **135** 49-56.

[Lawrence *et al.*, 1998] Lawrence,J.G., Ochman,H. (1998) Molecular archaeology of the Escherichia coli genome. *Proc. Natl. Acad. Sci. USA* **95** 9413-7.

[Maddox, 1998] Maddox J: What Remains to Be Discovered. TheFree Press, Simon & Schuster, New York. 1998.

[Martin, 1999] Martin W: Mosaic bacterial chromosomes: a challenge enroute to a tree of genomes. BioEssays 1999 21:99-104.

[Faguy and Doolittle, 2000] Faguy DM, Doolittle WF: Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. Trends Genet. 2000 May;16(5):196-7.

[Doolittle, 1999] Doolittle WF. Lateral genomics. Trends Cell Biol. 1999 Dec;9(12):M5-8.

[Brown and Doolittle, 1999] Brown JR, Doolittle WF. Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. J Mol Evol. 1999 Oct;49(4):485-95.

[Woese, 1998] Woese C. The universal ancestor. Proc Natl Acad Sci U S A. 1998 Jun 9;95(12):6854-9.

[Jain *et al.*, 1999] Jain et al. Horizontal gene transfer among genomes: The complexity hypothesis. PNAS 1999 96:3801.

[Moreire, 2000] Moreire: Multiple independent horizontal transferes of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. Mol Microbiol 2000 35:1.

[Forterre and Philippe, 1999] Forterre P, Philippe H. Where is the root of the universal tree of life? Bioessays. 1999 Oct;21(10):871-9.

[Nelson *et al.*, 1999] Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM, et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature. 1999 May 27;399(6734):323-9.

[Hou, 1999] Hou YM: Transfer RNAs and pathogenicity islands. Trends Biochem Sci. 1999 Aug;24(8):295-8.

[Hacker *et al.*, 1997] Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H: Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 1997 23(6):1089-1097.

[Alfano *et al.*, 2000] Alfano JR, Charkowski AO, Deng WL, Badel JL, Petnicki-Ocwieja T, van Dijk K, Collmer A: The pseudomonas syringae hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. Proc Natl Acad Sci U S A. 2000 Apr 25;97(9):4856-61.

[Groisman and Ochman, 1996] Groisman EA, Ochman H: Pathogenicity islands: bacterial evolution in quantum leaps. Cell. 1996 Nov 29;87(5):791-4.

[Kroll *et al.*, 1998] Kroll JS, Wilks KE, Farrant JL, Langford PR: Natural genetic exchange between Haemophilus and Neisseria: intergeneric transfer of chromosomal genes between major human pathogens. Proc Natl Acad Sci USA 1998 95(21):12381-5.

[Saunders *et al.*, 1999] Saunders NJ, Hood DW, Moxon ER: Bacterial evolution: bacteria play pass the gene. Curr Biol 1999 9(5):R180-3.

[Vasquez *et al.*, 1995] Vasquez JA, Berron S, O'Rourke M, Carpenter G, Feil E, Smith NH, Spratt BG: Interspecies recombination in nature: a meningococcus that has acquired a gonococcal PIB porin. Mol Microbiol 1995 15:1001-1007.

[Lawrence and Ochman, 1997] Lawrence JG, Ochman H: Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 1997 44(4): 383-397.

[Karlin *et al.*, 1998] Karlin S, Campbell AM, Mrazek J: Comparative DNA analysis across diverse genomes. Annual Rev Genet 1998 32:185-225.

[Josse *et al.*, 1961] Josse K, Kaiser AD, Kornberg A: Enzymatic synthesis of deoxyribonucleic acid VIII. Frequencies of neaarest neighbor base sequences in deoxyribonucleic acid. J. Biol Chem 1961 263: 864-875.

[Karlin *et al.*, 1994.a] Karlin S, Ladunga I, Blaisdell BE: Heterogeneity of genomes: measures and values. Proc Natl Acad Sci USA 1994 91: 12837-12841.

[Karlin *et al.*, 1994.b] Karlin S, Mocarski ES, Schachtel GA: Molecular evolution of herpesviruses: genomic and protein sequence comparisons. J Virol 1994 68: 1886-1902.

[Karlin *et al.*, 1994.c] Karlin S, Ladunga I: Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA 1994 20 91: 12832-12836.

[Kaldewaij, 1990] Kaldewaij A: Programming: The Derivation of Algorithms. Prentice Hall Int Series in Comp. Science, Prentice Hall, New York. 1990.

[Mirsky, 1999] Mirsky J: Genome Analysis Methodologies for the Identification of Horizontally Acquired DNA. Oxford University Computing Laboratory M.Sc. thesis. 1999.

[Burge *et al.*, 1992] Burge C, Campbell AM, Karlin S: Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA (1992) 89(4): 1358- 1362.

[Tettelin *et al.*, 2000] Herv Tettelin, Nigel J. Saunders, John Heidelberg, Alex C. Jeffries, Karen E. Nelson, Jonathan A. Eisen, Karen A. Ketchum, Derek W. Hood, John F. Peden, Robert J. Dodson, William C. Nelson, Michelle L. Gwinn, Robert DeBoy, Jeremy D. Peterson, Erin K. Hickey, Daniel H. Haft, Steven L. Salzberg, Owen White, Robert D. Fleischmann, Brian A. Dougherty, Tanya Mason, Anne Ciecko, Debbie S. Parksey, Eric Blair, Henry Cittone, Emily B. Clark, Matthew D. Cotton, Terry R. Utterback, Hoda Khouri, Haiying Qin, Jessica Vamathevan, John Gill, Vincenzo Scarlato, Vega Masignani, Mariagrazia Pizza, Guido Grandi, Li Sun, Hamilton O. Smith, Claire M. Fraser, E. Richard Moxon, Rino Rappuoli, and J. Craig Venter. The complete genome sequence of Neisseria meningitidis serogroup B strain MC58. Science (2000) 278: 1809-1815

[Karlin *et al.*, 1997] Karlin S, Mrazek J, Campbell AM: Compositional Biases of Bacterial Genomes and Evolutionary Implications. Journal of Bacteriology 1997 179(12): 3899-3913.