Uptake Signal Sequences in the Genome of *Haemophilus influenzae*

- An Interim Report -

T.Y. Chen[†], R.J. Redfield[⋆] and H.C. LEE[◇†1]

[◇]*Department of Life Sciences, National Central University, Chungli 320, Taiwan*

[†]*Department of Physics & Center for Complex Systems, National Central University, Chungli 320, Taiwan*

and

[⋆]*Department of Zoology, University of British Columbia, Vancouver, BC, Canada*

(Draft 2000 November 26)

The 1.8 Mbp genome of *Haemophilus influenzae* carries 1471 copies of its DNA uptake signal sequence (USS), a 29 bases long oligonucleotide including a 100% conserved 9-base core, AAGTGCGGT, 66% of the which are embedded in the 1738 known and putative genes of the genome. The frequency of occurrence of the 9-base (core) USS is about 100 times statistical expectation. Together with its flanks, the USS in *H. influenzae* take up 2.4% of the otherwise streamlined genome. The embedding of so many USS in the genome is a penalty to the organism that must be balanced by benefit to the organism derived from maintaining and evolving the uptake system. As a first step to understanding this balance we want to characterize and determine this cost. This is an interim report on that effort. In particular, we show that the USS is almost never embedded in the most conserved part of a gene.

[1]Send correspondence to: hclee@phy.ncu.edu.tw

INTRODUCTION

Genetic exchange is a major driving force in bacterial evolution, and is often assumed to have evolved for the purpose [1]. Many bacteria can take up DNA and thus acquire genes affecting virulence, host range and antibiotic resistance. DNA uptake by bacteria is one of the most primitive form of genetic exchange. Several important human pathogens *Haemophilus influenzae, Actinobacillus actinomycetemcomitans, Neisseria meningitidis and N. gonorrhoerae* have sequence-specific DNA uptake systems [2]. *Haemophilus* and *Neisseria* preferentially take up homologue DNA, by recognizing a short, highly-repeated sequence, the USS [2, 3, 4, 5]. The 1.8 Mbp genome of *H. influenzae* previously have been reported to have 1465 copies of a 29-base long USS [6, 7] (our count is 1471, see below). The full USS includes a 100% conserved 9-base core, AAGTGCGGT, and flanks including two 6-base oligonucleotides that are 85% conserved. The frequency of occurrence of the 9-base (core) USS is greater than 100 times statistical expectation. USS-dependent DNA uptake appears to be the best evidence that selection promotes genetic exchange [8, 9]. However we know of no selective processes that could have produced this system, and it is just as likely to have arisen from forces unconnected to genetic exchange. Once we understand its function, USS may also provide targets for intervention in infections by naturally-competent organisms - either as a drug-delivery mechanism or as a process to be inhibited by new drugs.

We observe that 66% of the USS (975 of 1471) reside in 38% of the known and putative genes (656 of 1738). Of the genes that carry USS, 433 has one USS, 152 has two, 56 has three, 8 has four, 6 has five and one gene (HI1685) has eight USS. Together with its flanks, the USS in *H. influenzae* take up 2.4% of the otherwise streamlined genome.

The presence of a USS in a DNA sequences coded for a gene implies a potential restriction of the functionality of the gene, hence imposes a cost to the adaptiveness of the organism. The embedding of so many USS in the genome therefore is a penalty to the organism that must be balanced by benefit to the organism derived from maintaining and evolving the uptake system. As a first step to understanding this balance we want to characterize and determine this cost. This is an interim report on that effort.

The genome of *H. influenzae* has 33 and 18 sections coded for transfer RNAs and ribosomal RNAs, respectively. None of these contain any USS. Possible significance of this absence will not be discussed here.

METHODS

Source. The complete *H. influenzae* genome is obtained from TIGR [10], NCBI Accession number L42023. The genome is 1,830,104 bp long, contains 1738 genes, 33 transfer RNAs and 18 ribosomal RNAs.

Overall strategy. Embedding a predetermined "alien" sequence - a USS in our case - in a DNA sequences coded for a gene could restrict the ability of the gene to function or to evolve, and the severity of the putative restriction would depend on the embedding position in the transcribed protein sequence. The severity would be high if the position contain residues crucial to the function or structure of the protein, and less high otherwise. The basic assumption we take is that the importance (to the structure and/or function of the protein) of a section in a protein sequence is positively correlated to how well that section is conserved in homologs of the protein. Here we quantify the restriction by examining the conservation of two kinds of sequences: the USS-embedded peptide (UEP) and the section of amino acids sequences that contain the UEP. Both cases require the searching of homologs of a gene that carry a (or more than one) USS.

Conservation of UEP sites. We examine the conservation of UEP's by the method described below. For simplicity we describe cases where the gene carries a single USS. (a) Take for a query sequence a 100 amino acids long with a UEP as nearly as possible and call this sequence the query. (b) Use a standard search software (e.g., BLAST) to find matches with E-values not exceeding E_0, and for each query select the three matches with the lowest E-values. We consider two cases, $E_0 = 1$ and $E_0 = 10^{-20}$. Queries with less than three qualified matches are discarded. (c) At this stage both the query and the matches include blank spaces inserted by the alignment process. Denote the query including insertions by Q, and a generic match by M. (d) Choose a match M and line it up with Q. Denote the UEP by q and the corresponding sequence in M by m. Depending on the relative position of the USS in the ORF of the host gene, the length of q is either three or four. (e) Use a scoring matrix (here we use "blosum62" [11]) to compute the similarity scores S_{qm} of q versus m and $S_{mm'}$ of m versus m', where m and m' are the "UEP"'s from different matches. Since for each query there are three matches, there are three S_{qm}'s and three $S_{mm'}$'s respectively. (f) A low score may be a non-conservation of the UEP or, especially in the case of $E_0 = 1$, due simply to the fact that Q and M are not homologs. In order to minimize this effect, we compute relative scores by normalizing S_{qm} by S_{QM} and $S_{mm'}$ by $S_{MM'}$, where S_{QM} and $S_{MM'}$ are the matching scores Q versus M and M versus M' (computed by "blast2seq"), respectively. (g) For each query, the means of the (normalized) S_{qm} and $S_{mm'}$ are taken to provide the coordinates of one point in a 2D-plot. Results are shown in Figs. 3 and 4 for $E_0 = 1$ and $E_0 = 10^{-20}$, respectively.

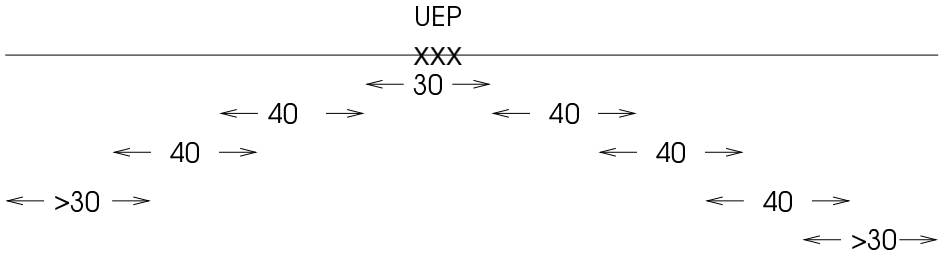
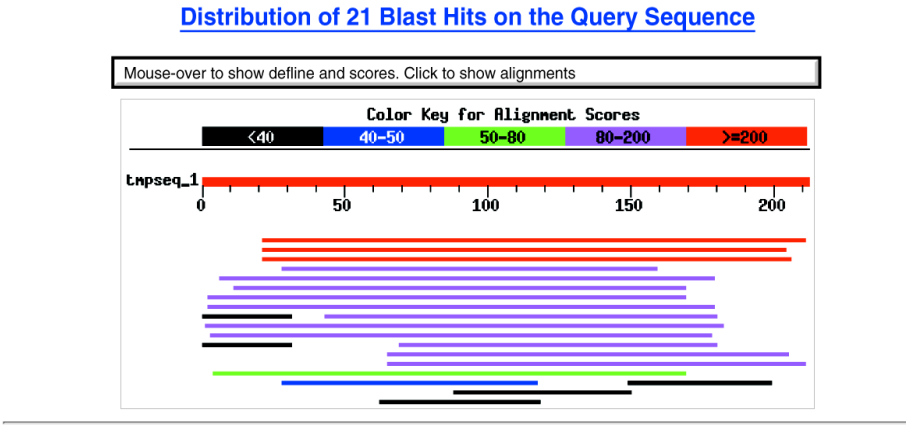


Figure 1: Segmentation of a protein sequence. XXX is the position of the UEP in the query, or of the corresponding peptides in a match.

Conservation of UEP containing segments. The idea is to see how well conserved is the segment of the gene (protein) sequence in which the UEP is embedded relative to other segments of the gene. Here only genes containing one USS are considered.
(a) Take the protein sequence of a gene containing a single USS as a query. (b) Use BLAST to search for matches which have not less than 40% indentities with the query and have lengths not shorter than 80% of the length of the query. Queries without qualified matches are discarded. (c) Trim the query and the matches, including insertions, to the length of the shortest match, and denote the trimmed query by Q, and a generic trimmed match by M. (d) Divide each (trimmed) sequence into several overlapping segments as follows. The segment containing the UEP is 30 peptides long. Other segments are 40 peptides long, except the end segments, which are back stepped from the end to make it at least 30 peptides long. Consecutive segments overlap by 10 peptides, except the overlaps involving the end segments, which may be longer. There is no overlap between the UEP and non-UEP segments. This is indicated in Fig. 1. (e) For a given match, a score for query versus match is computed with the score matrix "pam30" [12] for each segment, normalized by the query (segment) versus query score. This normalization is necessary because scores are length dependent.

If there are more than one match, this exercise is repeated for all matches and the scores, for each segment, are averaged over the matches. These scored will hereafter be called segment scores. The segment scores are then sorted according to magnitude and partitioned into four quartiles for analysis.

Automation. Computation and plotting, including interfacing with GenBank and its search engines, are automated using software written for this work.



Sequences producing significant alignments:		Score (bits)	E Value
sp|P44464|LIPB_HAEIN	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	400	e-111
sp|P30976|LIPB_ECOLI	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	255	5e-68
sp|Q9X6V9|LIPB_PSEAE	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	212	7e-55
sp|O75627|LIPB_HUMAN	PROBABLE LIPOATE-PROTEIN LIGASE B (LIP...	157	2e-38
sp|O32961|LIPB_MYCLE	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	119	4e-27
sp|Q9ZC91|LIPB_RICPR	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	112	5e-25
sp|Q9X6X4|LIPB_MYXXA	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	106	5e-23
sp|Q10404|LIPB_MYCTU	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	103	3e-22
sp|O36017|LIPB_SCHPO	PUTATIVE LIPOATE-PROTEIN LIGASE B (LIP...	93	5e-19
sp|P74519|LIPB_SYNY3	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	90	3e-18
sp|O23021|LIPB_ARATH	PROBABLE LIPOATE-PROTEIN LIGASE B (LIP...	89	8e-18
sp|Q00520|LIPB_PARVE	PROBABLE LIPOATE-PROTEIN LIGASE B (LIP...	88	2e-17
sp|Q06005|LIPB_YEAST	PROBABLE LIPOATE-PROTEIN LIGASE B, MI...	85	1e-16
sp|O13476|LIPB_KLULA	LIPOATE-PROTEIN LIGASE B, MITOCHONDRIA...	83	3e-16
sp|O19898|LIPB_CYACA	PROBABLE LIPOATE-PROTEIN LIGASE B (LIP...	77	3e-14
sp|Q51854|LIPB_PROHO	LIPOATE-PROTEIN LIGASE B (LIPOATE BIOS...	44	2e-04
sp|P09432|NTRC_RHOCA	NITROGEN REGULATION PROTEIN NTRC	31	2.3
sp|P13848|VG7_BPPH2	HEAD MORPHOGENESIS PROTEIN (LATE PROTEI...	30	5.1
sp|O05974|DP3A_RICPR	DNA POLYMERASE III, ALPHA CHAIN	29	6.6
sp|P07533|VG7_BPPZA	HEAD MORPHOGENESIS PROTEIN (LATE PROTEI...	29	6.6
sp|Q9Y9E6|LIPB_AERPE	PROBABLE LIPOATE-PROTEIN LIGASE B (LIP...	29	6.6

Figure 2: BLAST hits for the query sequence of the protein lipB encoded by the gene HI0027.

RESULTS AND DISCUSSIONS

Sample search result. Fig. 2 shows a typical BALST search result, where the query is the protein sequence of the lipoate biosynthesis protein B (lipB) encoded by the gene HI0027. The first hit, or match, is the query itself. The last two columns in the bottom half of the

figure give the scores and E-values of the matches, respectively. Not shown are the query-match alignments and the number of identities in each alignment. In this example, the first eight matches have E-values less than 10^{-20} and all are homologs in different organisms of the query. Only the second and third hits have lengths not less than 80% of the query. Hence for the study UEP containing segments only these are kept as matches and the other hits are discarded. In this case, the length of the original query is 212 but the trimmed length is 180.

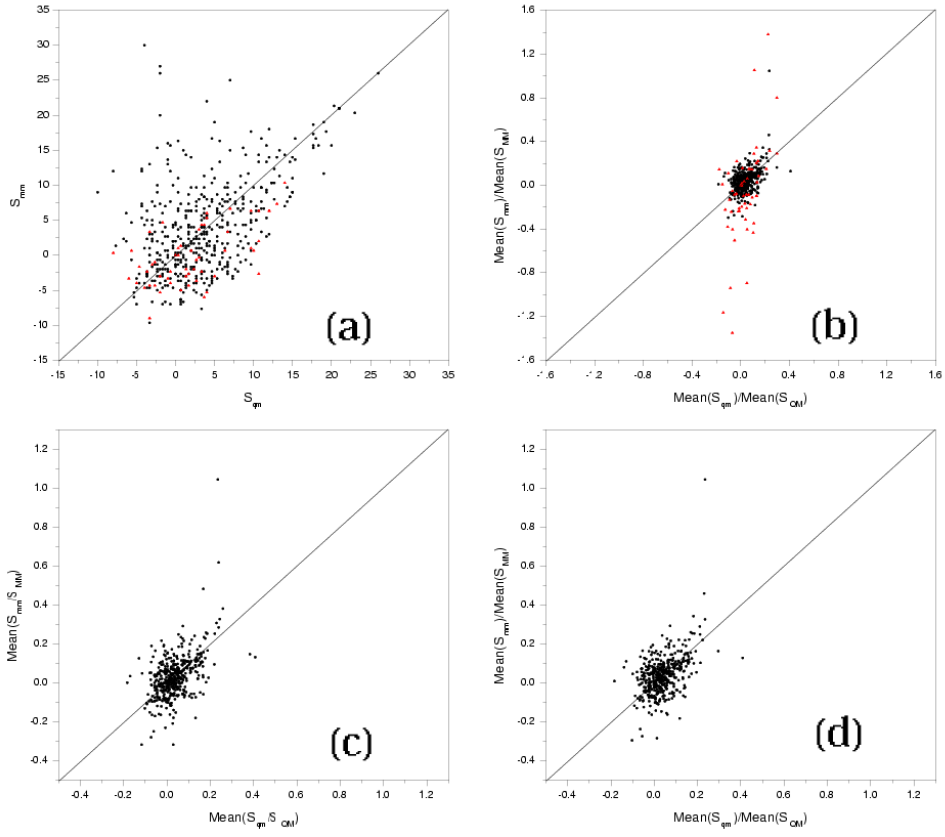


Figure 3: Conservation of UEP sites. (a) S_{qm} *vs* S_{mm}. Each point represents a case with one query and three matches. The ordinate is the mean of three match versus match UEP scores and the abscissa is the mean of three query versus match UEP scores. Total number of points is 514. The 56 red point indicate cases when the BLAST score for at least on of the S_{MM}'s is zero. (b) Same as (a), except that $(S_{qm})_{ave}/(S_{QM})_{ave}$ *vs* $(S_{mm})_{ave}/(S_{MM})_{ave}$ are plotted. Total number of points is 514. (c) Same as (a), except that $(S_{qm}/S_{QM})_{ave}$ *vs* $(S_{mm}/S_{MM})_{ave}$ are plotted. Total number of points is 458; the "red" case in (a) and (b) are excluded. (d) Same as (b) with the red points removed.

Composition of UEP. The three-peptide TAL appears 618 times in the protein sequences of *H. influenzae* and of these 369, or 59.71%, are encoded by USS. These latter constitute 63.6%, or 369 out of 580, of the total number three-peptide UEP's. The three-peptide SAV

appears 426 times in *H. influenzae* and of these 269, or 63.15%, are encoded by USS. These 269 UEP's constitute 68.1% of the total number of 395 four-peptide UEP's.

Result on coservatin of UEP sites: UEP sites are only slightly disrupted by USS. Figs. 3 and 4 show the results of analysis of UEP conservation. In Figs. 3 the cut-off E-value is 1 while in Figs. 4 the cut-off E-value is 10^{-20}. In the figures the subscript "Q" denotes a USS embedded query sequence; "M" denotes a match sequence; "q" denotes the USS-encoded peptides (UEP, 3 or 4 amino acids) in the query sequence; "m" denotes the corresponding peptides in a match sequence. In the plots in the two figures, each point gives the result for one UEP.

When the cut-off E-value is 1, there are 514 cases of UEP queries with at least three matches. The mean of S_{qm} (averaged over matches) against the mean of $S_{mm'}$ for these 514 cases are plotted in (a) of Fig. 3. In Table 1 more detailed information of the data shown in the plot is given: the average $\bar{S}$ of all $|S_{mm'}|$ and $|S_{qm}|$ is 5.32; the number of cases out 514 when $S_{mm'} > S_{qm}$ is 213; the number of when $S_{mm'} = S_{qm}$ is 23; the average $\Delta_>$ of $(S_{mm'} - S_{qm})/\bar{S}$ over the set $S_{mm'} \geq S_{qm}$ is 0.92; the average $\Delta_<$ of $(S_{mm'} - S_{qm})/\bar{S}$ over the set $S_{mm'} \leq S_{qm}$ is -0.73; the average of $(S_{mm'} - S_{qm})/\bar{S}$ over all data is -0.0058.

Table 1: Result on comparison USS-encoded peptides with corresponding peptides a match sequences. Notation: q stands for S_{qm}; m for S_{mm}; Q for S_{QM}; M for S_{MM}; prime on q, m Q or M indicates averaged over matches; N = total number of cases in figure; $N_>$ = data in the set $y \geq x$; $\bar{S}$ = average of all $|x|$ and $|y|$; $\Delta_>$ = average of $(y-x)/\bar{S}$ over set $y \geq x$; $\Delta_<$ = average of $(y-x)/\bar{S}$ over set $y \leq x$; Σ = average of $(y-x)/\bar{S}$ over all data.

Cut-off E-value	Test (x *vs* y)	Figure	N	$N_>$[†]	$\bar{S}$	$\Delta_>$	$\Delta_<$	Σ
1	q *vs* m	3 (a)	514	213+23	5.32	0.92	-0.73	-0.0058
	q'/Q' *vs* m'/M'	3 (d)	458	215+4	0.067	0.99	-0.92	-0.016
	$(q/Q)'$ *vs* $(m/M)'$	3 (c)	458	195+7	0.069	1.02	-0.90	-0.064
10^{-20}	q *vs* m	4 (a)	91	41+11	8.08	0.69	-0.37	0.19
	q'/Q' *vs* m'/M'	4 (b)	91	45+1	0.058	0.73	-0.47	0.13
	$(q/Q)'$ *vs* $(m/M)'$	4 (c)	91	41+2	0.056	0.78	-0.47	0.11

[†] Number after the plus sign is the number of data with y=x.

With the cut-off E-value for matches being as large as 1, the taxonomic distances between the query and match sequences may vary widely, and the normalization of the UEP scores ($S_{mm'}$ and S_{qm}, respectively) by the sequence scores ($S_{MM'}$ and S_{QM}, respectively) should reduce the dependence of the UEP scores on the varying distances. In Figs. 3, (b) plots the mean of S_{qm} divided by mean of S_{QM} (the x coordinate) against the mean of $S_{mm'}$ divided by mean of $S_{MM'}$ (the y coordinate). The red points are the 56 cases in which at least one (but not all three) of the match versus match scores given by BLAST is zero. Figs. 3 (d) shows the same points in (b) except the red points. In (c) the 56 case giving the red points in (b) are excluded and the mean of S_{qm}/S_{QM} (the x coordinate) are plotted against the mean of $S_{mm'}/S_{MM'}$ (the y coordinate). Numerical analyses of data shown in (c) and (d) are given in Table 1.

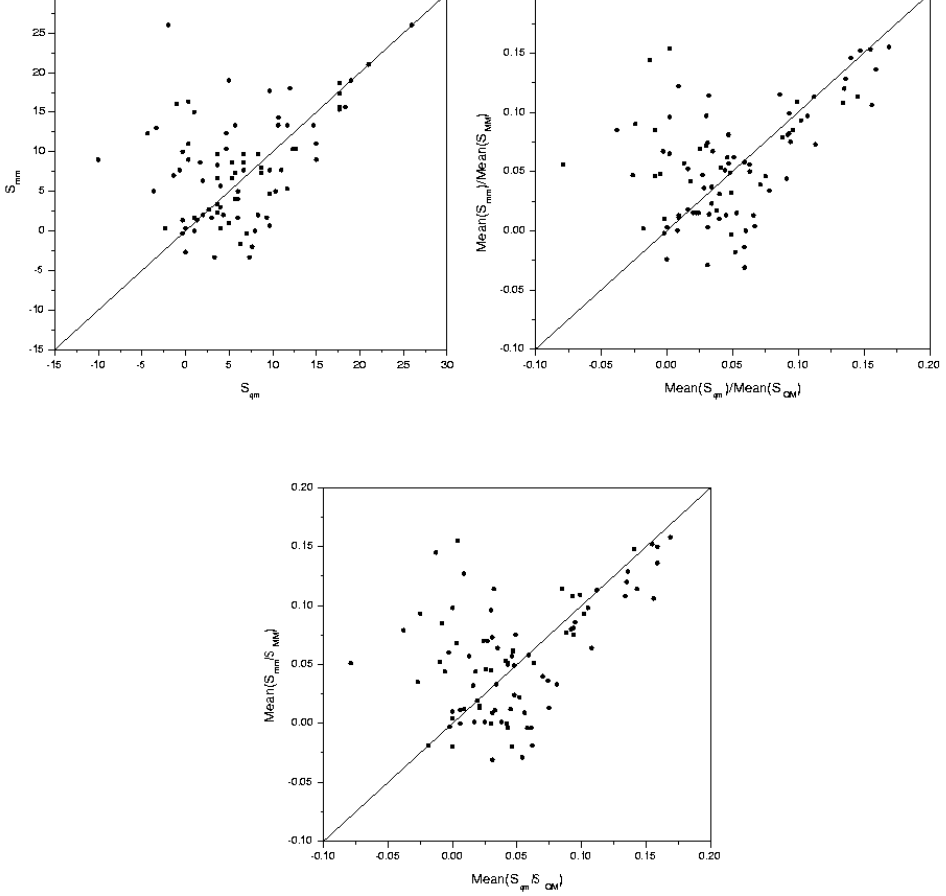


Figure 4: Same as Fig. 3, except that matches were selected with cut-off E-value of 10^{-20}. Total number of points is 91. There is no case involving a vanishing S_{MM}. (a) S_{qm} *vs* S_{mm}. (b) $(S_{qm})_{ave}/(S_{QM})_{ave}$ *vs* $(S_{mm})_{ave}/(S_{MM})_{ave}$. (c) $(S_{qm}/S_{QM})_{ave}$ *vs* $(S_{mm}/S_{MM})_{ave}$.

These results show the following trend. There are fewer cases having $S_{mm'} > S_{qm}$ than the opposite, but for such cases the average difference between $S_{mm'}$ and S_{qm} is larger, such that there is almost complete cancellation in the average Σ of $(S_{mm'} - S_{qm})/\bar{S}$ over all cases. Σ is expected to be of the order of unity if there were no cancellation, but is instead -0.0058 in case (a). Renormalizing the UEP score by the sequence gives a noticeable effect: the magnitude of Σ in case (c), while still small, is about ten times greater than its value in case (a). We conclude that a typical UEP in a host protein sequence differs little from from peptides in sequences homologous to the host at locations corresponding to those of the UEP in the host sequence.

When the cut-off E-value is 10^{-20}, there are only 91 cases of UEP queries with at least three matches. In thess cases the query and all three match sequences are always homologs and none of the match versus match score ($S_{MM'}$) is zero. The plots in Fig. 4 display data similar to those shown in (a), (b) and (c) of Fig. 3, but for the new set of 91 cases of queries and matches: (a) plots S_{qm} *vs* S_{mm}; (b) plots $(S_{qm})_{ave}/(S_{QM})_{ave}$ *vs* $(S_{mm})_{ave}/(S_{MM})_{ave}$, where there are no red points because all $S_{MM'}$ are nonzero; (c) plots $(S_{qm}/S_{QM})_{ave}$ *vs* $(S_{mm}/S_{MM})_{ave}$. Numerical analyses of data shown in Fig. 3 are given in the lower part of

Table 1.

Now the numbers of cases haivng $S_{mm'} > S_{qm}$ and the opposite are very close to being equal, but the average difference when $S_{mm'} > S_{qm}$, $\Delta_>$, is decisively greater than its counterpart $\Delta_<$ when the opposite is true. The un-normalized $\Sigma = 0.19$ is still much less than unity, but is alarmingly more than a factor of 30 greater in magnitude than Σ when the E-value is 1. Normalizing the UEP score by sequence score reduces Σ to 0.11, which is now less than a factor of two greater than by a factor greater than Σ when the E-value is 1.

Notwithstanding the smallness of the magnitude of Σ, the changing of the its sign from negative to positive when the cut-off E-value is changed from 1 to 10^{-20} has a reasonable interpretation. When the cut-off E-value is larger, the chances are greater that the mathces are taxinomically unrelated, this produces a bias to reduce the similarity scores between matches (*i.e.*, $S_{mm'}$) relative to the scores between query and match (*i.e.*, S_{qm}). This produces a net effect that pushes δ towards the negative side. We therefore expect δ at a E-value of 10^{-20} to be greater than δ at a E-value of 1, as is seen in Table 1.

The preceeding analysis gives us confidence in taking the positive sign of δ when the E-value is 10^{-20} as being significant. It means that in sequences homologous to the protein (query) sequence carrying UEP, peptides at locations corresponding to those occupied by the UEP are more similar among the matches then they are between query and match. In other word, the UEP does seem to have intruded upon its host, and have prodcued a detectable disruption.

Result on conservation of UEP containing segments: UEP are seldom found in the most conserved parts of proteins. The set of segment scores for the peptide sequence of the gene HI0027, which encodes the lipoate biosynthesis protein B (lipB), is shown in Fig. 5.

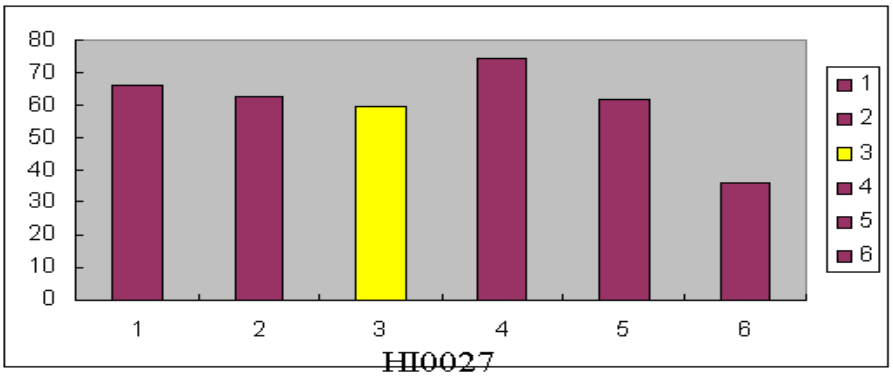


Figure 5: Relative segment similarity scores for the USS containing gene HI0027. Query is the peptide sequence of the protein lipB encoded by the gene. For each segment scores are averaged over the match segments and normalized by the score of the query segment against itself divided by 100. See text for more detail.

The ordinate gives relative (query) segment versus match scores, averaged over the matches - there are two in this case, normalized by the segment versus segment score and multiplied by 100. The scores may be roughly interperpreted as percentage similarity. In the

figure, the towers represent segments along the sequence. The position of the yellow tower, which represents the UEP containing segment, indicates that the UEP is located near the center of the query sequece. The scores for the UEP containing segment is 60, which lies between the low value of 36 and the high value of 85. The differences between these values are significant. For a rough feeling of what these score could mean consider the following. If for simplicity we divide the twenty kinds of amino acids into four equivalent sets of five, then the probabilities (roughly, the E-value) that two 40-peptide sequences are 85% and 36% similar are 3.4×10^{-21} and 2.1×10^{-9}, respectively, and the probability that two 30-peptide sequences are 60% similar is 1.4×10^{-11}.

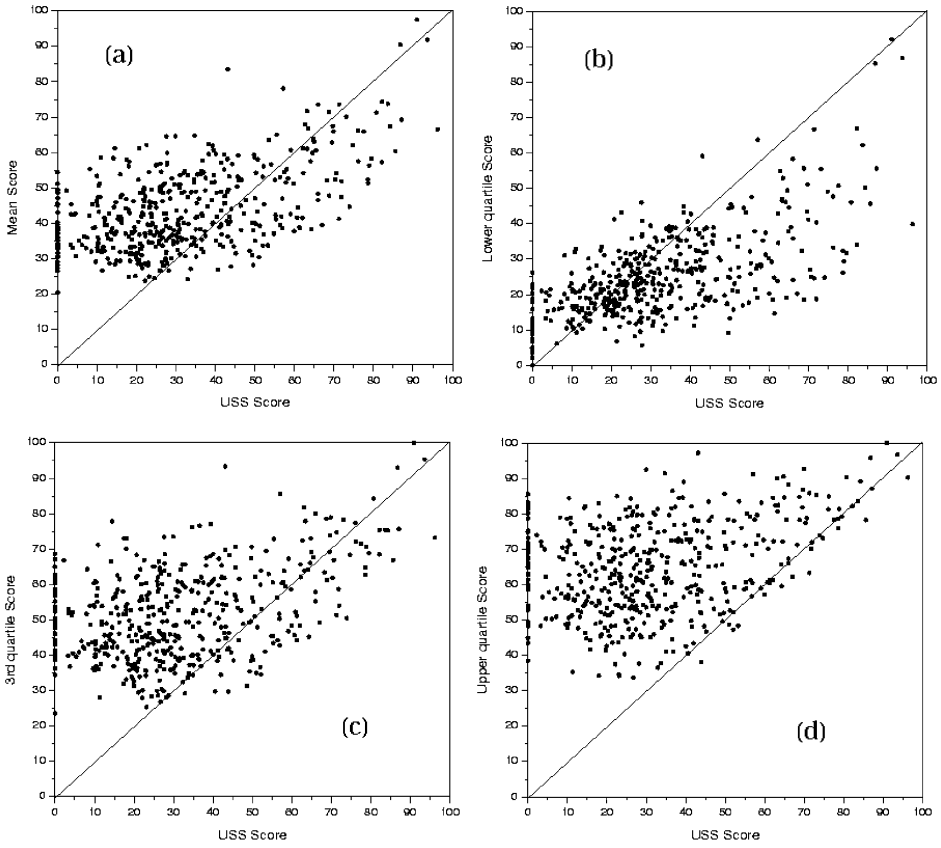


Figure 6: Relative scores for sectors in proteins containing UEP. Each plot has 473 point representing the protein sequences of 473 USS carrying genes. The abscissa of a point is the score for the UEP containing segment and the ordinate of the point is (a) the averaged score of all sectors, (b) the averaged score for the lower quartile, (c) the averaged score for the third quartile and (d) the averaged score for the upper quartile.

Fig. 5 shows that the number of segments is six, which is not a multiple of four. In such a case, which is true for most genes, the segments are partitioned (according to magnitude) into four quartiles proximity in magnitude. For example, for the case at hand, segment 6 is partitioned into the first or lower quartile; segment 3, the UEP segment, is partitioned into the second quartile; segments 2 and 5 are partitioned into the third quartile; segments 1 and

4 are partitioned into the fourth or upper quartile. Cases with less than four segments are discarded.

There are 473 USS containing genes that meet the criteria for analysis. Of these genes 86 of the proteins encoded by them are conserved hypothetical and 24 are putative; the others are known. Six pieces of data are extracted from each analyzed gene: the averaged score over all segments, and averaged score for each quartile, and the score for the UEP containing segment. The UEP containing segments do not distribute evenly in the four quartiles: 271, 101, 51 and 50 are in the lower, second, third and upper quartiles, respectively.

Fig. 6 shows some important aspects of these results. Each of the four plot has 473 points and every point represents the protein sequence of a USS carrying gene. The abscissa of a point is the score for the UEP containing segment and the ordinate of the point is (a) the averaged score of all sectors (b) the averaged score for the lower quartile, (c) the averaged score for the third quartile and (d) the averaged score for the upper quartile, respectively. It is seen that the score for the UEP containing segment is more than likely higher than the average score in the lower quartile, but more than likely lower than the average score in the third quartile, and almost never higher than the average score in the upper quartile. In the figures, points lying on the y-axis are those points for which "pam30" gives no score for the UEP containing segment.

*** Discuss the few UEP' in highly conserved regions ***

The conclusion is that on average, UEP are embedded in less conserved half of the protein sequence. They are almost never embedded in the upper most conserved quartile of the protein sequence. In other words, UEP are highly selective in where they reside in a protein.

This result is consistent with what was seem earlier about the conservation of the UEP itself. In most cases the UEP in a protein are only slightly conspicuous when they are compared with oligopeptides at corresponding locations in homologous proteins, because those locations lie in relatively less conserved regions in the proteins.

Summary. USS' in *H. influenzae* are not randomly distributed over its whole gonome. They show a marked favor for coded regions: 66% of the USS are in 38% of genome coded for genes. The USS are approximately evenly distributed over the coded regions, but they display a pronounced bias in avoiding the most conserved segments of the genes they do reside in. The USS are only slightly conspicuous in the genes: the USS-encoded peptides - the UEP - are only slightly noticeable in the protein encoded by the gene when compared to peptides in corresponding sites in homologous proteins.

We do not know how the USS were generated in the first place. If they were randomly generated over the whole genome then, over time, we expected the USS in the intergenic region to erode more rapidly, simply because they are not protected from evolution by the genes. This should be one of the factors contributing to coded regions being favored in the distribution of USS. That USS' mostly reside in less conserved segments of genes could also be caused by evolution pressure: a USS placed in the most conserved segment - by whatever mechanism that generates USS' - would be more likely to disrupt the function of the protein coded by the host gene, and would be less likely to have itself established. Since UEP containing segments in protein sequences are rarely highly conserved, UEP sites themselves should not be expected to be highly conserved. This is consistent with the observation that UEP' are only slightly conspicuous among peptide occupying similar sites in other proteins. Phenotypically, one may say that avoiding the most conserved segments in a gene is how

USS minimizes the restriction of the functionality of the gene.

A description of the origin and evolution of USS' has not yet been given. When it is, the bias in the distributions of USS' in the genome of *H. influenzae* and of UEP sites within USS embeded genes should be among the useful clues.

References

[1] A. Kondrashov. "Classification of hypotheses on the advantage of amphimixis. J. Hered. **84** (1993) 372-87.

[2] S.H. Goodgal, "DNA uptake in *Haemophilus* transformation", Ann. Rev. Gen. **16** (1982) 169-92.

[3] D.Danner, H.O. Smith and S. Narang, "Construction of DNA recognition sites active in *Haemophilus* transformation", PNAS **79** (1982) 2393-7.

[4] S.D. Goodman and J.J. Scocca, "Identification and arrangement of DNA sequence recognized in specific transformation of *N. gonorrhoerae*", PNAS **85** (1988) 6982-6.

[5] S.H. Goodgal, "Sequence and uptake specificity of cloned sonicated fragnebts of *H. influenzae* DNA", J. Bact. **172** (1990) 5924-8.

[6] H.O. Smith, J.-F. Tomb, B. Dougherty, R. Fleischmann and J. Venter, "Frequency and distribution of DNA uptake signal sequences in the *H. influenzae* Rd genome", Science **269** (1995) 538-40.

[7] S. Karlin, J. Mrazek and A.M. Campbell, "Frequent oligonucleotides and peptides of the *H. influenzae* genome", Nucleic Acid Res. **24** (1996) 4263-72.

[8] M.G. Lorenz and W. Wacknagel, "Bacterial gene transfer by natural genetic transformation in the environment," Micro. Rev. **58** (1994) 563-602.

[9] J.S. Kroll, K.E. Wilks, J.L. Farrant and P.R. Langford, "Natural genetic exchange between *Haemophilus* and *Neisseria*: intergenic transfer of chromosomal genes between major human pathogens." PNAS **95** (1998) 12381-5.

[10] http://www.tigr.org/tigr-scripts/CMR/mol_info.spl?db=ghi

[11] S. Heinikoff and J.G. Heinikoff, "Position based sequence weights". PNAS (USA) **89** (1992) 10915-10919.

[12] M.O. Dayhoff, R.M. Schwartz and B.C. Orcutt, " A model of evolutionary change in proteins". In *Atlas of Protein Sequence and Structure*, M.O. Dayhoff, ed. (National Biomedical Research Foundation, Washington, DC., 1978) pp 345-362.