doi:10.1006/jmbi.2000.3576 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 297, 335–353

JMB

Inferring Regulatory Elements from a Whole Genome. An Analysis of *Helicobacter pylori* σ^{80} Family of Promoter Signals

Anne Vanet[1,2], Laurent Marsan[3], Agnès Labigne[2] and Marie-France Sagot[4]*

[1]*Institut de Biologie Physico-Chimique, UPR CNRS 9073, 13 rue Pierre et Marie Curie 75005, Paris, France*

[2]*Unité de Pathogénie Bactérienne des Muqueuses Institut Pasteur, 28, rue du Dr Roux, 75724, Paris, Cedex 15 France*

[3]*Institut Gaspard Monge, Cité Descartes, 5 boulvard Descartes, Champs-sur-Marne Marne-la-Vallée, France*

[4]*Service d'Informatique Scientifique, Institut Pasteur 28 rue du Dr Roux, 75724 Paris, Cedex, France*

**Corresponding author*

Helicobacter pylori is adapted to life in a unique niche, the gastric epithelium of primates. Its promoters may therefore be different from those of other bacteria. Here, we determine motifs possibly involved in the recognition of such promoter sequences by the RNA polymerase using a new motif identification method. An important feature of this method is that the motifs are sought with the least possible assumptions about what they may look like. The method starts by considering the whole genome of *H. pylori* and attempts to infer directly from it a description for a family of promoters. Thus, this approach differs from searching for such promoters with a previously established description. The two algorithms are based on the idea of inferring motifs by flexibly comparing words in the sequences with an external object, instead of between themselves. The first algorithm infers single motifs, the second a combination of two motifs separated from one another by strictly defined, sterically constrained distances. Besides independently finding motifs known to be present in other bacteria, such as the Shine-Dalgarno sequence and the TATA-box, this approach suggests the existence in *H. pylori* of a new, combined motif, TTAAGC, followed optimally 21 bp downstream by TATAAT. Between these two motifs, there is in some cases another, TTTTAA or, less frequently, a repetition of TTAAGC separated optimally from the TATA-box by 12 bp. The combined motif TTAAGC$\times(21 \pm 2)$-TATAAT is present with no errors immediately upstream from the only two copies of the ribosomal 23 S-5 S RNA genes in *H. pylori*, and with one error upstream from the only two copies of the ribosomal 16 S RNA genes. The operons of both ribosomal RNA molecules are strongly expressed, representing an encouraging sign of the pertinence of the motifs found by the algorithms. In 25 cases out of a possible 30, the combined motif is found with no more than three substitutions immediately upstream from ribosomal proteins, or operons containing a ribosomal protein. This is roughly the same frequency of occurrence as for TTGACA$\times$(15-19)TATAAT (with the same maximum number of substitutions allowed) described as being the σ^{70} promoter sequence consensus in *Bacillus subtilis* and *Escherichia coli*. The frequency of occurrence of the new motif obtained, TTAAGC$\times$(19-23)TATAAT, remains high when all protein genes in *H. pylori* are considered, as is the case for the TTGACA$\times$(15-19)TATAAT motif in *B. subtilis* but not in *E. coli*.

Keywords: combined motif; description inference; promoter; *Helicobacter pylori*; prokaryotes

Introduction

Helicobacter pylori is a Gram-negative, spiral-shaped pathogenic bacterium. It specifically colonizes the gastric epithelium of primates and is the

Abbreviations used: CDS, coding sequence; EF, elongation factor; EM, Expectation Maximization; RBS, ribosome binding site.

E-mail address of the corresponding author: sagot@pasteur.fr

0022-2836/00/020335–19 $35.00/0

etiologic agent of chronic gastritis (Blaser, 1992). The properties of the bacterium associated with host and various environmental factors can cause gastritis to progress, over a period of years, to more severe diseases. Such diseases include peptic ulcer, gastric lymphoma, gastric atrophy and carcinoma (Correa, 1995; McColl, 1996).

The RNA polymerase σ factor of *H. pylori*, related to σ^{70}, is a σ^{80} polymerase, and its 4.2 region differs substantially from that of other bacteria, in particular *Escherichia coli*. The RNA polymerase itself is β-β′ fusioned. These observations suggest that *H. pylori* promoters may also be different. Furthermore, *H. pylori* operons are known to be difficult to express in *E. coli* (Beier *et al*., 1998). We therefore tried to identify motifs possibly involved in the recognition of *H. pylori* promoter sequences by the RNA polymerase.

The whole genome of *H. pylori* (containing 1,667,867 bp) was completely sequenced at TIGR and the result published in July 1997 (Tomb *et al*., 1997). The set of sequences we extracted from the genome corresponds to all the non-coding regions on both strands located upstream from the genes (according to TIGR's annotation). The algorithm presented here attempts to infer a description for a family of promoters directly from the chosen set of sequences. This approach is different from a search for promoters by similarity with a previously established description (based, for instance, on some biological experimental evidence frequently acquired by analysing other organisms) (Chen *et al*., 1995; Helmann, 1995; Stormo, 1990a,b).

Previous purely algorithmical methods for inferring conserved motifs, such as promoters, from a set of sequences fall into two main categories. The first comprises the statistical approaches that try to locate DNA-binding sites by maximizing the value of some function of the quantity of information present in the site (Bailey & Elkan, 1995; Baldi *et al*., 1994; Cardon & Stormo, 1992; Crowley *et al*., 1997; Krogh *et al*., 1994; Lawrence *et al*., 1993; Lawrence & Reilly, 1990; Stormo, 1990a,b; Stormo & Hartzell, 1989). This is in some way related to the quantity of conservation in the site. Due to the complexity of the problem and the size of the search space to be explored, none of these methods is exhaustive. They are often also sensitive to the presence of excessive noise in the data.

The second category of algorithms are motif-based methods. The oldest algorithms proceed by generating all possible words of length w for some reasonable value of w (Galas *et al*., 1985; Mengeritsky & Smith, 1987; Queen *et al*., 1982, Waterman, 1984); they become impractical when searching for longer motifs (typically of length greater than ten base-pairs), or a combination of small ones. Such algorithms frequently make the assumption that the studied sequences have been pre-aligned on the basis of some biological criterium, for instance, the start of transcription (Galas *et al*., 1985). None of the more recent algorithms allows for errors, and all may therefore find only perfectly conserved motifs (Ulyanov & Stormo, 1995; van Helden *et al*., 1998). Some authorize a small number either of wild-cards, e.g. of positions in the sites where any base is permitted, or of degenerate letters (e.g. purine, pyrimidine, etc.) (Fraenkel *et al*., 1995; Ulyanov & Stormo, 1995; Worlfertstetter *et al*., 1996). One algorithm (Fraenkel *et al*., 1995) also allows for spacers, that is, for a variable number of wild-cards inside a motif. The total length of sequence such spacers may cover is, however, limited. It cannot, for instance, bridge the optimal distance of 15 to 19 bp that separate the two promoter sequences recognized by the σ^{70} family of some bacterial RNA polymerases. These sequences are believed to be present around positions −10 and −35 upstream from the start of transcription (Record *et al*., 1996).

The promoters in prokaryotes are, in general, known to be a combination of at least two cooperating motifs separated from one another by strictly defined, sterically constrained distances. However, there is no method available, other than the heuristical EM-based approach (EM, Expectation Maximization) of Cardon & Stormo (1992) to infer a combination of motifs in a single pass of the algorithm from a set of unaligned, potentially noisy sequences. We propose an algorithm which allows such inferences to be made.

Here, we also suggest a combined motif for the promoter sequences involved in the binding of the σ^{80} family of RNA polymerases in *H. pylori*. A preliminary analysis of the genome performed at TIGR has shown (Tomb *et al*., 1997) that the bacterium has possibly only three types of σ factors, and therefore of promoter families: σ^{80}, σ^{54} and σ^{28}. The first family is the most frequently observed in all well-studied bacteria. The motif we propose for the σ^{80} family is supported by some simple statistical considerations and by the fact that it is present upstream from genes which are in general very well expressed. It is also in agreement with the experimental results independently obtained by Wosten *et al*. (1998b) in a related organism, *Campilobacter jejuni*.

Results

Overview of the approach

Two different algorithms were used to infer conserved motifs in a set of unaligned sequences. The two are based on the idea of inferring motifs by comparing words in the sequences to an external object. This object is a word over the same alphabet as that of the sequences that satisfies the following general property: it must be present with at most e errors in at least a percentage q of the sequences of the set, where e and q are user-defined values. Parameter q is called a quorum; the only errors authorized here were substitutions. The first algorithm, called 1, identifies single motifs. The second, called 2, permits the inference of a combination of two motifs separated from one

another by a range of distances. Motifs of the second type model pairs of sites recognized by a same protein and which, therefore, must stand at a strict distance from one another. Algorithms 1 and 2 were run on three sets of sequences: (i) set A which corresponds to all regions non-coding on both strands located upstream from genes; (ii) set B, which is a subset of A and contains the regions located between divergent genes; and, finally, (iii) set C which consists of all regions upstream from ribosomal RNA genes, or genes encoding ribosomal proteins, alone or as part of an operon. The number of sequences and of nucleotides in each set is given in Table 1. Further details concerning the data and algorithmic approach may be found in Materials and Methods.

Validating the method on test sets

Since the algorithms used are novel, we started by validating them on two test sets. The sets are composed of well-established sequences from *E. coli* and *B. subtilis* containing an experimentally determined transcription start or, sometimes, promoter. The sets were obtained from Ozoline *et al*. (1998) and Helmann (1995), respectively (see Materials and Methods).

Algorithm 1 was run to extract single motifs of length six or more base-pairs with one substitution allowed for quorums of 5% (*E. coli*) and 50% (*B. subtilis*). The results are presented in Table 2. Some motifs appear to be variants of a single motif, both in terms of their labels and of their occurrences (most correspond to approximately the same words in the sequences). The motifs shown in Table 2 are grouped into families; the grouping was done manually.

Pairs of motifs were identified using algorithm 2. To check that the distance between the two parts of a pair is specific to most promoter sequences (as suggested in the literature for both bacteria (Record *et al*., 1996)), we looked for conserved pairs separated by a distance, d plus or minus 1, where d varied between 9 and 23. Since combined motifs are longer, the quorums were fixed at a lower value than for single motifs with one substitution allowed: 2% for *E. coli* and 10% for *B. subtilis*. The results are given in Figure 1.

In both *E. coli* and *B. subtilis*, the statistically most significant motifs (using a χ^2 test, see Materials and Methods) are TATAAT at −10 and TTGACA at −35, separately or as a pair. Interestingly, in *E. coli*, the motifs were identified at very low quorums only. Higher values of q yielded no significant single or combined motifs (that is, no motifs with probability of appearing by chance below 10^{-3}). In the case of combined motifs, none was found at a quorum of 4%.

Applying the method to *H. pylori*

Looking for single motifs

Algorithm 1 on set A. Algorithm 1 was run to extract single motifs from the 756 sequences of set A with two different kinds of parameters (0 and 1 substitution for quorums of 25 and 50%, respectively). The results obtained are presented in Table 3. Motifs appear manually grouped into families and, inside each family, classified by their statistical significance, as determined by a χ^2 test (see Materials and Methods). Only those motifs with a chance probability below 10^{-5} (meaning a χ^2 value of 19.5 or more) are reported in Table 3. Motifs with probability between 10^{-3} (χ^2 value of 10.8 or more) and 10^{-5} are submotifs or shifted versions of those presented in Table 3 (results not shown). Six families were obtained and give rise to various comments.

The first is that a shadow of the TATA-box is found among the statistically most significant motifs, in some cases, an exact TATAAT motif (262 occurrences with no error, χ^2 of 81.16). The CTAAAA motif included in this family (found with no error) may correspond not to the TATA-box, but to the −35 box of the σ^{28} in *B. subtilis*.

The positions of the occurrences of many of the motifs of the second and third families represent some of the base-pairs in the so-called Shine-Dalgarno sequence. This is especially true of the occurrences of motifs from the third family (some of the motifs of the second family may correspond to a sequence other than the Shine-Dalgarno). Figure 2 shows the positional distribution for the occurrences (with up to one substitution) of one of the most frequent and statistically significant motifs in family 3, AAGGAG (only the 50 positions upstream from the start of translation are indicated). Indeed, the Shine-Dalgarno sequence is usually located five to ten base-pairs upstream from the translational start codon on the mRNA,

Table 1. Number and total length of sequences in sets A, B and C for *H. pylori, B. subtilis* and *E. coli*

		Set A	Set B	Set C
H. pylori	Total number of sequences	756	340	30
	Total number of nucleotides	168,709	97,360	7066
E. coli	Total number of sequences	2652	1194	37
	Total number of nucleotides	596,355	335,648	9078
B. subtilis	Total number of sequences	2721	1076	41
	Total number of nucleotides	560,898	326,470	20,459

Table 2. Simple motifs found with Algorithm 1 in *E. coli* and *B. subtilis* test sets

	Escherichia coli quorum 5% one substitution				*Bacillus subtilis* quorum 50% one substitution			
Family 1	ATAATGCGG	34	4	25	TATAATA	94	49	31
	TATAATGCGC	23	2	19	GTATAAT	74	36	23
	ATAATGCGC	30	6	17	TGTTATA	66	36	14
	TGTGTATA	47	16	17	ATAATAT	82	52	14
	ACAATGCGC	24	4	15	ACAATA	108	82	13
	AAAATGCGC	29	6	15	TAAAATA	95	68	12
	ATGATGCGC	24	4	15	TATTATA	76	48	12
	TATCATGC	40	13	15	GATATA	98	71	12
	GTATAATGC	24	4	14	TATAGT	95	69	11
	TAATGCGG	41	14	14				
	GGTATACT	31	8	14				
	CTATAATGC	23	4	14				
	TATACTGA	38	13	13				
	GGTAGAAT	34	11	13				
	TGGTAGAA	33	10	13				
	CTGTATAA	42	16	13				
	GTGTATAA	42	16	13				
	AATGCGCG	40	15	13				
	TATAATGCT	25	6	12				
	TGTATAAT	51	23	12				
	TAATGCGC	45	19	12				
	CTGTTTAT	46	19	12				
Family 2	GTTGACAC	36	11	14	TTTTACA	76	47	13
	CTGATAGA	31	8	14	GTGACA	68	39	13
	TCACACTT	36	11	14	GTTGAC	66	39	12
	TGACACTT	38	12	14	TTTACAA	75	48	11
	ACACTTAT	41	15	13				
	GCTGACA	64	32	12				
Family 3	AAAACAGT	52	21	15				
	AAAAACAGT	28	7	13				
Family 4	TTACGCTG	39	13	14				
	TTACGCAT	43	17	12				
	TGTTACGC	39	14	12				
	TTTACGCT	44	18	12				
	TACGCTG	66	34	12				
Family 5	CTGAAAAA	53	24	12				
Family 6	GTTACACT	34	11	12				
Family 7	TAACTCTG	32	10	12				
Family 8	AAAGCGCC	35	12	12				
Family 9	AACCTGAA	36	13	12				

Column 1 corresponds to motifs of length greater than six base-pairs identified, with at most one substitution and a quorum of 5% or 50%, column 2 to the number of sequences where at least one occurrence of the motif was found, column 3 to the same for shuffled versions of the sets (average over 1000 simulations) and, finally, column 4 to the χ^2 value. All motifs with a probability of happening by chance below 10^{-3}, and only those are shown, up to a maximum of 40.

and binds with specificity to the 3′ sequence of the 16 S RNA molecules in prokaryotes (it is thus also called the RBS, for ribosome binding site). This 3′ sequence is the same in *H. pylori* as it is in *E. coli*. It was therefore expected that the motif related to the Shine-Dalgarno sequence in both bacteria would be very similar. This motif is, in general, described as being a substring of AGGAGGTGA (Record *et al.*, 1996).

The fourth family seems to correspond to a novel motif and will be further examined below.

In some cases, motifs of the fifth family occur close to one another and may form the stem of a stem-loop structure. One such structure, with

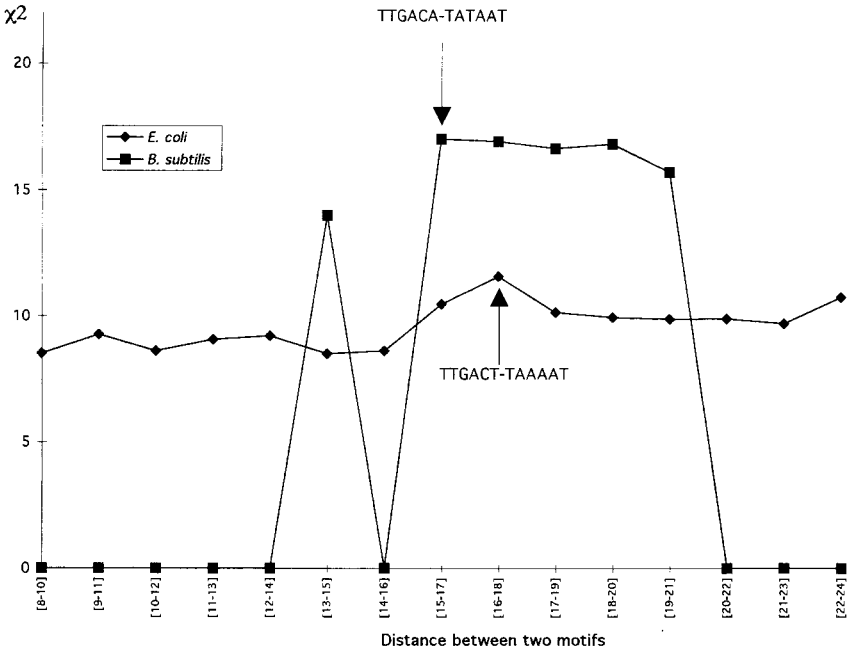

Figure 1. Statistical significance of the observed peak in the distribution of distances between the two elements of the combined motifs found by algorithm 2 in the test sets for *E. coli* and *B. subtilis* with up to one substitution allowed and a quorum of 2% for *E. coli* and 10% for *B. subtilis*. For this, a χ^2 and a *Z*-score were calculated (only χ^2 values are shown) between the number of occurrences observed (allowing for the same maximum number of subtitutions) in all sets against that observed on average in shuffled versions of each. The most significant motifs identified for each interval are shown above the curves (only for the intervals located at or near a peak).

almost the same stem sequences as one of the motifs here (ACGCT twice), is possibly involved in the promoter activity of a *Campilobacter coli* gene (Kinsella *et al*., 1997). Further investigation is needed to assess the relevance of these motifs to gene expression in *H. pylori*.

Finally, no function can at the present time be proposed for motifs in the sixth family.

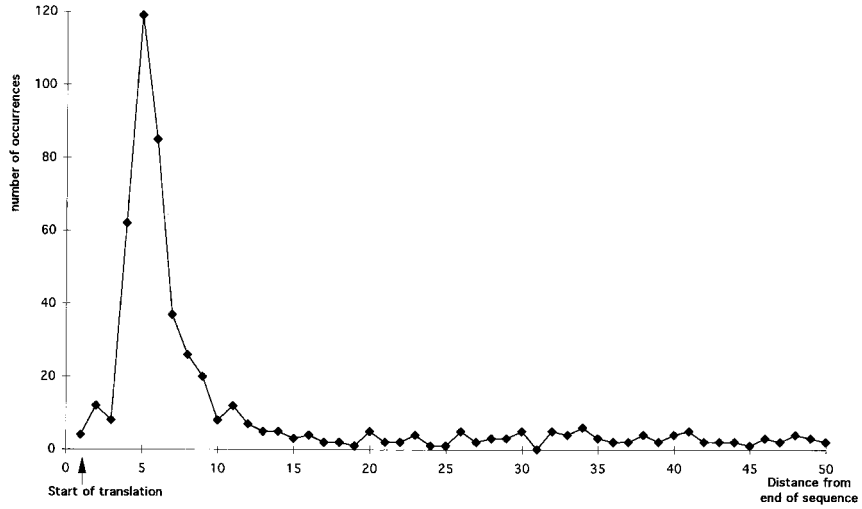

Figure 2. Distribution of the positions with respect to the start of translation of the occurrences of one of the most frequent and statistically significant motifs, AAGGAG (Shine-Dalgarno), found by algorithm 1 in set A with a quorum of 50% and up to one substitution allowed. Only the 50 positions before the start of translation are shown. Positions more than 50 bases upstream from this start have a flat, close to zero, distribution (not shown).

Table 3. Simple motifs found with algorithm 1 in set A for *H. pylori*

	Quorum 25% no substitution				Quorum 50% one substitution			
Family 1	TATAAT	262	111	81	TATAATC	455	268	93
	TAAAAT	358	246	34	TATGATA	407	226	89
	TATAA	413	301	33	TATCAT	603	448	75
	ATTATA	191	111	26	TATAATA	543	380	74
	CTAAAA	243	160	23	CTATAAT	455	288	74
	ATAAT	380	292	21	TATCATA	386	223	73
	TAAAAC	196	124	21	TTATAAT	602	456	67
					TATAATG	441	283	66
					TATGAT	593	455	59
					ATGATA	593	458	57
					ATAATAC	400	255	57
					TAGAATA	453	312	53
					GTTATAA	475	335	52
					TAAAACC	457	318	51
					TAGAAT	664	555	50
					TAAAATC	552	421	50
					GTATAAT	400	264	50
					TACAATA	418	282	49
					ATTATAA	582	456	49
					TATTATA	513	381	48
					GCTATAA	384	253	47
					TACAAT	640	530	46
Family 2	TAAGG	304	187	41	TAAGGG	573	420	68
	TTTAAG	288	177	38	TTAAGGG	414	259	64
	TTAAGG	195	100	38	TAAGGAT	391	256	49
	TTTTAG	265	170	29	TTAAGGA	460	326	47
	TTAAG	409	306	28				
Family 3	AAGGA	365	198	79	AAGGAG	576	408	82
	AAAGG	365	269	25	AAAGGAG	407	255	62
					AGGGGT	423	286	50
					AAGGAT	619	499	50
					AAGGATA	394	263	47
Family 4	TTTTA	638	528	46	GGTTTTA	488	338	60
	TTAAAA	472	343	45	GGGTTTT	404	256	59
	TTTTAA	476	347	44	GGGGTT	434	290	55
	TAAAA	623	520	38	GATTTTA	569	435	53
	TTTAAAA	301	197	32	GATTTTAA	406	270	50
	ATTTTA	357	249	32	GTTTTAA	605	486	47
	GGGTT	214	124	31	ATTTTAAG	416	286	45
	GAGTT	205	117	31				
	GTTTT	441	338	28				
	TTTAA	612	524	27				
	GTTTTT	286	197	24				
	TTTTAAA	287	199	24				
	ATTTTAA	215	137	22				
	GTTTTA	206	130	22				
	TTTTTA	438	351	20				
	GATTTT	225	150	20				
Family 5	AACGC	192	77	59	AGCGCT	408	275	47
	GCGTT	196	83	56	GCGTTTT	400	267	48
Family 6	ATCAA	281	187	27				
	AAAATC	213	138	21				

Column 1 corresponds to motifs of length greater than five base-pairs identified (with no error and a quorum of 25%) or six (with at most one substitution and a quorum of 50%), column 2 to the number of sequences where at least one occurrence of the motif was found, column 3 to the same for shuffled versions of the sets (average over 1000 simulations) and, finally, column 4 to the χ^2 values. All motifs with a probability of happening by chance below 10^{-5}, and only those are shown, up to a maximum of 40.

Algorithm 1 on set B. The same parameters were used to look for single motifs which are statistically significant in the sequences of set B. The results are given in Table 4. Only the 40 most statistically significant motifs are shown.

These results generally confirm and strengthen those obtained by inferring motifs from set A. The motif from family 6 (described above), ATCAAA, was not found to be significant in set B (using a χ^2 statistic with one degree of freedom) at a

Table 4. Simple motifs found with algorithm 1 in set B for *Helicobacter pylori*

	Quorum 25% no substitution				Quorum 50% one substitution			
Family 1	TATAAT	148	64	49	TATCATA	243	121	88
	ATTATA	147	64	48	TATGATA	243	122	86
	TAAAAT	196	122	33	TATCAT	316	226	73
	TAAAA	305	252	28	TATTATA	292	192	71
	ATAATA	118	61	25	TATAATA	292	194	70
	TTATA	218	155	23	ATGATA	316	230	69
	TATAA	220	160	22	TATAATC	246	143	64
	TATTAT	117	63	22	GATTATA	245	143	62
	ATAAT	209	150	20	TTATAAT	309	225	61
					ATTATAA	309	226	60
					CTATAAT	249	150	60
					ATAATAC	233	133	59
					GTATTAT	233	133	59
					ATTATAG	248	149	59
					TATTGTA	238	142	54
					TATTATAA	217	121	54
					TACAATA	238	143	53
					TTATAATA	217	122	53
					TATTCTA	246	154	51
					TAGAATA	247	156	51
					ATTATAAT	211	119	50
					ATCATA	304	229	49
					TATAAT	210	120	48
					TTATTATA	208	119	47
					TATGAT	304	232	45
					ATTATAC	220	134	43
					TATACTA	203	118	43
					TATAATG	231	146	43
					TTATAAC	251	168	43
					CATTATA	230	148	40
					GTTATAA	251	171	40
					TAGTATA	203	122	39
					GTATAAT	221	140	39
					TAAAATCA	192	112	38
					TTATCAT	229	149	38
					TAAAATC	281	209	38
Family 2	TCCTT	168	90	38	TTTTAAGC	177	114	24
	CCTTA	155	87	30	TTTAAGC	229	185	12
	TAAGG	156	89	29				
	CTTAAA	146	83	26	TTTTAAGG	186	106	39
	TTAAGG	102	47	26				
	CCTTAA	102	47	26				
	TTTAAG	145	83	26				
	CTTAAAA	97	44	25				
	TTTTAAG	96	43	25				
Family 3	AAGGA		170	94				
Family 4	ATTTTA	197	121	34	GATTTTA	282	206	42
	TTTTA	305	246	34	TGATTTTA	192	108	42
	TTTTAA	228	164	24	ATTTTAG	283	212	37
	TTAAAA	228	165	24				
Family 5	AACGC	86	40	21				

Column 1 corresponds to motifs of length greater than five base-pairs (with no error and a quorum of 25%) or six (with at most one substitution and a quorum of 50%), column 2 to the number of sequences where at least one occurrence of the motif was found, column 3 to the same for shuffled versions of the sets (average over 1000 simulations) and, finally, column 4 to the χ^2 value. All motifs with a probability of happening by chance below 10^{-5}, and only those are shown, up to a maximum of 40 (plus the two (T)TTTAAGC motifs).

probability below 10^{-5} with either choice of parameters (results not shown). Only one seemingly new motif appeared when up to one substitution was allowed. This is motif (T)TTTAAGC (χ^2 value of 23.65) located in family 2 of Table 4. This is very similar to the (T)TTAAG(G) motifs identified either in set A, or when no error was permitted.

Setting the quorum at various other ranges (between 15% and 30% for no error, and between 30% and 60% for up to one substitution) revealed

Table 5. Simple motifs found with algorithm 1 in set C for *H. pylori*

	Quorum 25% no substitution				Quorum 50% one substitution			
Family 1	TATAAT	15	6	6	TATTTAA	30	22	9
	AAAAT	27	20	5	TAGTATAA	18	7	8
	CAAAATC	19	10	6	ATCTAAAA	21	10	8
					AAAATCT	26	17	7
					TACAATC	16	6	7
					TCTATAAT	15	6	6
					TCAAAAGA	17	7	6
					AAAATC	30	24	6
					TCTAAAA	27	19	6
					GAAAAT	30	25	6
Family 2	TTTAAG	18	6	7	TTTAAGCG	15	5	8
	TTTAAGC	9	1	7	TTTTAAGC	21	11	7
	AGCTAAA	8	1	5	TTTAAGA	28	20	7
	GCTAAA	10	3	5	GTTTAATC	16	6	7
	GCATG	8	2	5	TTAAGCAT	16	7	6
					CTTAAG	25	14	9
					TTTAAGGG	17	6	9
					TAAGGG	27	17	8
					TAAGGGG	17	7	7
					ATTTTGAG	17	8	6
					TAAGGGA	18	8	6
					CTTTTAAG	19	10	6
					ATCAAGG	15	6	6
					TCTTAAG	21	12	6
					TTTTAAGG	20	11	6
Family 3	AAGGA	19	7	10	CAAACGA	16	5	9
	AGGAGA	8	1	7	TCAAGGA	18	7	9
	GGAGA	10	3	5	AAGGAGA	19	8	9
	AAGGAG	8	1	5	CAAGGA	25	14	9
	AATGG	13	5	5	AAGGATAA	16	5	9
					AAACGAG	19	8	8
					AGGAGAT	17	6	8
					TAAGGAG	19	8	8
					CAAGGAT	17	7	8
					AAGGGT	26	17	7
					AGAGTTT	24	14	7
					AAGGACA	15	5	7
					GGGGTT	23	13	7
					AGGGGTT	18	8	6
					AAGGGG	23	14	6
Family 4	ATTAA	9	19	6	ATTATTTTA	19	9	6
	ATTTTA	22	12	6	CTATTTTA	20	11	6
	TTTAAAAAG	8	1	6	TGTAAAA	26	18	6
	TTTAAAAA	14	5	6	ATTTTTAGT	16	7	6
Family 5	ATCAA	15	6	5				
Family 6					AATCCCT	16	6	7

Column 1 corresponds to motifs of length greater than five base-pairs identified (with no error and a quorum of 25%) or six (with at most one substitution and a quorum of 50%), column 2 to the number of sequences where at least one occurrence of the motif was found, column 3 to the same for shuffled versions of the sets (average over 1000 simulations) and, finally, column 4 to the χ^2 values. All motifs with a probability of happening by chance below 5×10^{-2}, and only those are shown, up to a maximum of 40 (plus the motifs having TTAAGC at their core).

the same main families of motifs (not shown) as listed in Tables 3 and 4, in particular families 1 to 4, thus reinforcing their robustness. At a higher quorum, motifs from families 5 and 6 were sometimes missing (result not shown). Motif (T)TTTAAGC (Table 3) in particular disappeared when the quorum was 60% for up to one substitution. It nevertheless has a high χ^2 value.

Two of the other apparently robust motif families (4 and 5 in Table 4) are novel (that is, not previously described). At this point in the analysis, it was premature to assign with confidence any of

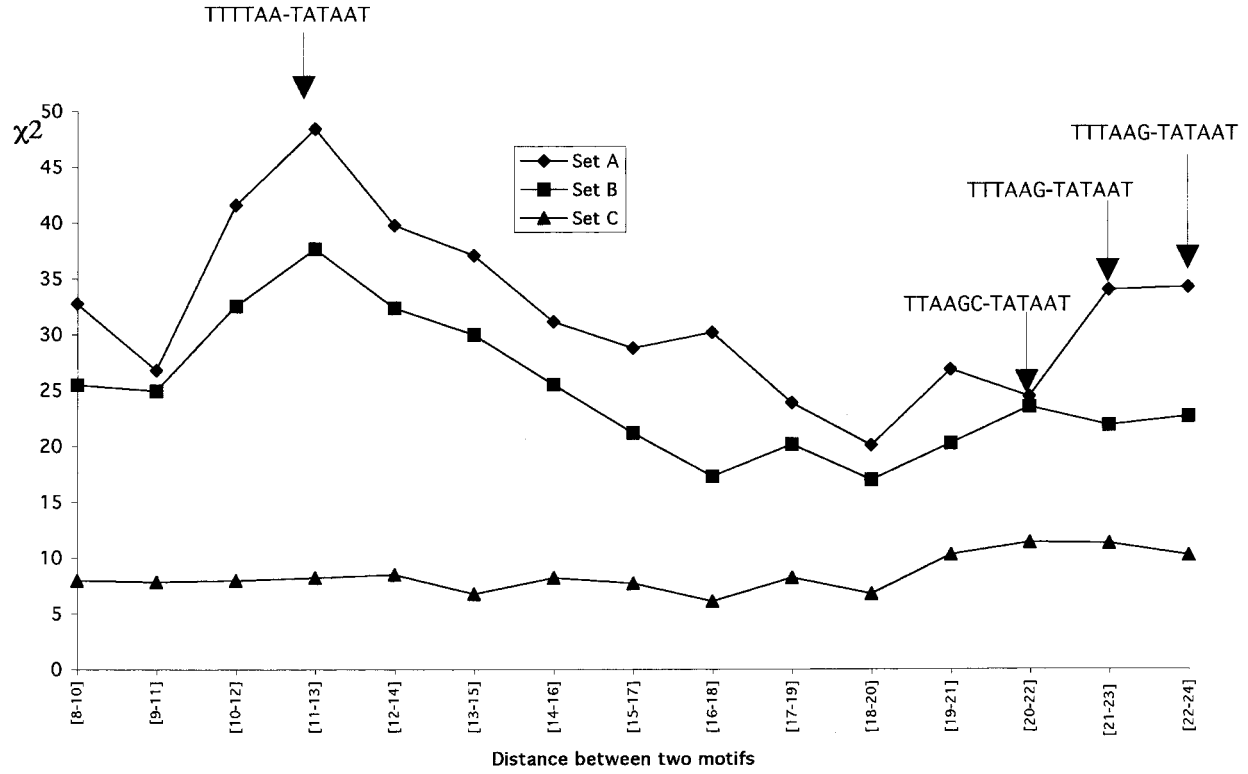

Figure 3. Statistical significance of the observed peak in the distribution of distances between the two elements of the combined motifs found by algorithm 2 in sets A, B and C of *H. pylori* with a quorum of 10% and up to one substitution allowed. For this, a χ^2 value and a *Z*-score were calculated (only χ^2 values are shown) between the number of occurrences observed in all three sets, against that observed (allowing for the same maximum number of substitutions) on average in shuffled versions of each (1000 simulations were performed). The most pertinent motif identified for each interval is shown above the curves (only for the intervals located at or near a peak).

these families to binding bacterial promoter sites (such as the one at −35).

Algorithm 1 on set C. Algorithm 1 was run on set C. This produced the results shown in Table 5. The χ^2 values are much lower than those obtained for sets A and B and we are near the limits of the applicability of a χ^2 statistic due to the small number of occurrences. The results were, however, confirmed by *Z*-scores (values above 10, not shown). These results do not contradict those obtained with sets A and B. Table 5 shows a new motif corresponding to family 6, AATCCCT, when up to one substitution is allowed. The motif could simply refer to a sequence complementary to a Shine-Dalgarno.

Looking for combined motifs

Algorithm 2 on sets A, B and C. Algorithm 2 was run to extract from sets A, B and C pairs of motifs separated by an user-defined interval. We first looked in all three sets for conserved pairs of motifs (overall, one substitution only allowed) separated by various intervals of distances, each motif of length at least six base-pairs. The intervals ranged from (8-10) to (22-24) by increments of 1. The quorum was fixed at 10% in all cases. Figure 3 shows the statistical significance of the motifs found for the various ranges, (8-10), (9-11), ... (22-24). Two peaks of unequal height are observed, one situated around 12 bp and the other around 21 bp. The motifs with highest statistical significance found at and near each peak are indicated in Figure 3. Only intervals with motifs having *Z*-scores (not shown in the Figure) above 10 are included.

Running algorithm 2 on set C more clearly defined the motifs significant to the two peaks, particularly the second one. Figure 3 shows that only one motif is significant at 10^{-3} (χ^2 value of 10.8 or more) within the interval (20-22). This corresponds to motif TTAAGC followed by TATAAT (χ^2 value of 11.34). The motif corresponding to the first peak identified in sets A and B (TTTTAA followed by TATAAT, interval 11-13) remains statistically significant in set C, but less than TTAAGC×(20-22)TATAAT.

All motifs found by algorithm 1 appear identified again by algorithm 2. This is particularly true of a TATA-box look-alike preceded by two main types of motifs: one that consists of a run of T nucleotides (in general three or four) followed by

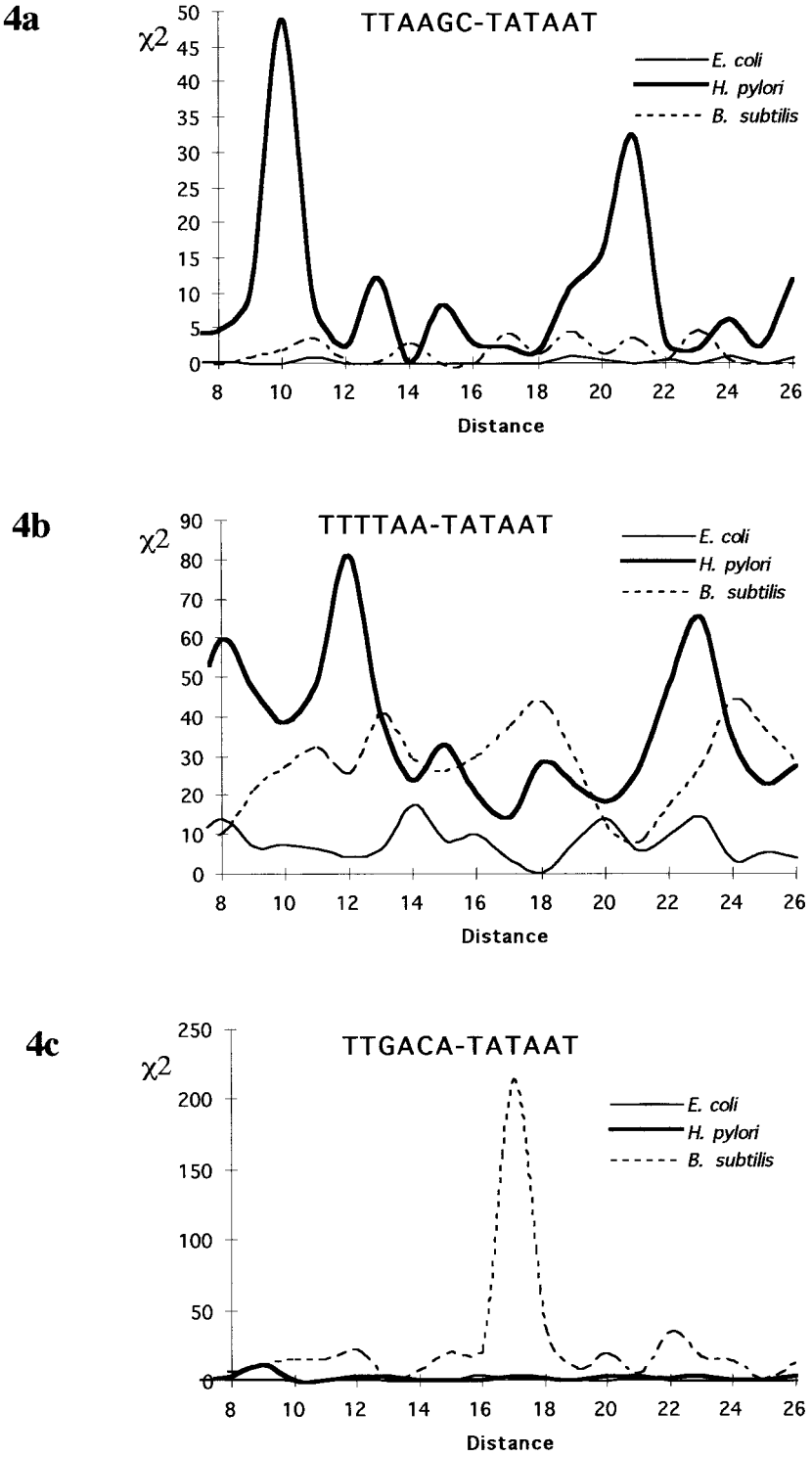

Figure 4. Statistical significance of the observed peak in the distribution of distances between the two elements of the combined motifs found in *H. pylori* (a) TTAAGC×(19-23)TATAAT, (b) TTTTAA×(10-14)TATAAT and (c) the classical consensus for *E. coli* and *B. subtilis*, TTGACA×(15-19)TATAAT. The χ^2 tests were performed between the number of occurrences observed (allowing for up to two substitutions) in set A against that observed in shuffled versions of set A (1000 simulations were performed), in all three organisms and for a range of distances larger than the one observed here or described in the literature. All curves shown have been locally smoothed in the following way: coordinate i on the x-axis represents in fact the average of the χ^2 values for distances i and $i+1$.

a run of A nucleotides (in roughly the same numbers); and the other that we may represent as (T)TTAAG(C/G).

Confirming the results obtained with algorithm 2 using set A and a comparative test with B. subtilis *and* E. coli. The two combined motifs identified with the highest statistical significance in Set C and potentially recognized by the σ^{80} factor of *H. pylori* were TTTTAA×(11-13)TATAAT and TTAAGC×(20-22)TATAAT.

Combined motifs are composed of essentially two elements: a pair of motifs and the distance between the motifs. Each element in the pair of the above motifs, TATAAT and TTTTAA, more strongly than TTAAGC, appears to be significant (see Tables 3, 4 and 5). To confirm the statistical significance of the combined motifs, we tried to: (i)

Table 6. Number and frequency of occurrence per sequence as well as statistical significance in sets A, B and C for *H. pylori*, *B. subtilis* and *E. coli* of the two main combined motifs identified by algorithm 2 in *H. pylori* and of the classical consensus

Subs.	Motif	*H. pylori* Set A	*H. pylori* Set B	*H. pylori* Set C	*E. coli* Set A	*E. coli* Set B	*E. coli* Set C	*B. subtilis* Set A	*B. subtilis* Set B	*B. subtilis* Set C
0	TTAAGC x(19-23) TATAAT	6 0.79 % 1.18 %	4 1.18 % 13.33 %	4 13.33 % 0	0 0% 0%	0 0% 0%	0 0% 0%	1 0.04 % 0%	0 0% 0%	0 0% 0%
	TTTTAA x(10-14) TATAAT	17 2.25 % 2.65 %	9 2.65 % 6.67 %	2 6.67 % 0	0 0% 0%	0 0% 0%	0 0% 0%	11 0.40 % 0.56 %	6 0.56 % 0%	0 0% 0%
	TTGACA x(15-19) TATAAT	3 0.40 % 0.59 %	2 0.59 % 6.67 %	2 6.67 % 0	1 0.04 % 0.08 %	1 0.08 % 0%	0 0% 0%	14 0.51 % 0.28 %	3 0.28 % 4.87 %	2 4.87 % 0
1	TTAAGC x(19-23) TATAAT	50 6.61 % 10.29 %	35 10.29 % 33.33 %	10 33.33 % 0	13 0.49 % 1.09 %	13 1.09 % 0%	0 0% 0%	13 0.48 % 0.56 %	6 0.56 % 0%	0 0% 0%
	TTTTAA x(10-14) TATAAT	143 18.92 % 18.24 %	62 18.24 % 36.67 %	11 36.67 % 0	42 1.58 % 1.93 %	23 1.93 % 2.70 %	1 2.70 % 0	131 0.81 % 6.78 %	73 6.78 % 0%	0 0% 0%
	TTGACA x(15-19) TATAAT	15 1.98 % 2.06 %	7 2.06 % 6.67 %	2 6.67 % 0	20 0.75 % 1.01 %	12 1.01 % 13.51 %	5 13.51 % 0	143 5.26 % 5.29 %	58 5.29 % 1.49 %	16 1.49 % 0
2	TTAAGC x(19-23) TATAAT	237 31.35 % **49.96** **32.84**	133 39.12 % **32.84** **4.83**	16 53.33 % **4.83** 0.14	233 8.79 % 0.14 0.37	133 11.14 % 0.37 1.14	4 10.81 % 1.14 0.98	249 9.15 % 0.98 1.79	143 8.38 % 1.79 0.02	4 9.76 % 0.02 0
	TTTTAA x(10-14) TATAAT	449 59.39 % **82.39** **74.30**	227 66.76 % **74.30** **6.68**	22 73.33 % **6.68** 0.24	420 15.84 % **5.75** 1.48	233 19.51 % 1.48 0.87	6 16.22 % 0.87 0.01	782 28.74 % **32.88** **54.94**	406 23.89 % **54.94** 0.24	9 21.95 % 0.24 0
	TTGACA x(15-19) TATAAT	116 15.34 % **6.55** 3.70	66 19.41 % 3.70 0.01	4 1.18 % 0.01 0	247 9.31 % 2.29 0.88	134 11.22 % 0.88 **7.93**	11 29.73 % **7.93** 0.01	634 23.30 % **162.76** **93.25**	304 17.82 % **93.25** **17.55**	23 56.10 % **17.55** 0.01

The motifs were searched with zero, one and two substitutions. Bold, underscored values correspond to a χ^2 value whose probability of appearing by chance is less than 10^{-5}, bold values to a χ^2 value whose probability of appearing by chance is less than 5×10^{-2}. The statistical significance is indicated in the case of two substitutions only.

check the statistical significance of the motifs TTTTAA×(11-13)TATAAT and TTAAGC×(20-22)TATAAT in sets A, B and C, this time permitting up to two substitutions and larger intervals of distance between the two parts of each combined motif (Table 6); (ii) check the statistical significance of the observed peak in the distribution of distances between the two elements of the motifs found by doing a χ^2 test (with one degree of freedom) between the number of occurrences observed (allowing for up to two substitutions) in all three sets and that observed in a shuffled version of the sets, for a range of distances larger than those observed above or known from the literature (Figure 4); (iii) check the ''optimality'' of the base present at each position of the motifs (i.e. how ''strong'' is each base in the inferred consensus for the promoter sequences) (Figure 5).

We conducted the same verifications with the motif generally described as being the σ^{70} promoter sequence in *E. coli* and *B. subtilis*, TTGACA×(15-19)TATAAT, and compared the results. For this purpose, we built sets of sequences A, B and C from the *E. coli* and *B. subtilis* genomes (the genomic sequences were retrieved from http://mol.genes.nig.ac.jp/ecoli/ and ftp://ncbi.nlm.nih.gov/genbank/ ~ genomes/bacteria/Bsub/, respectively) in exactly the same way as from *H. pylori* (see Table 1 for the number of sequences and of bases obtained in each case).

Table 6 and Figures 4 and 5 show that the only combined motifs that are statistically significant (χ^2 value with a probability below 10^{-5} of being due to chance alone) for both frequency of occurrence and optimality of the distance between the two parts of the motifs are: TTTTAA×(11-13)TATAAT and TTAAGC×(20-22)TATAAT in *H. pylori*, and TTGACA×(15-19)TATAAT in *B. subtilis*.

For *H. pylori*, the peaks in the distribution of distances between the occurrences in sets A, B or C of

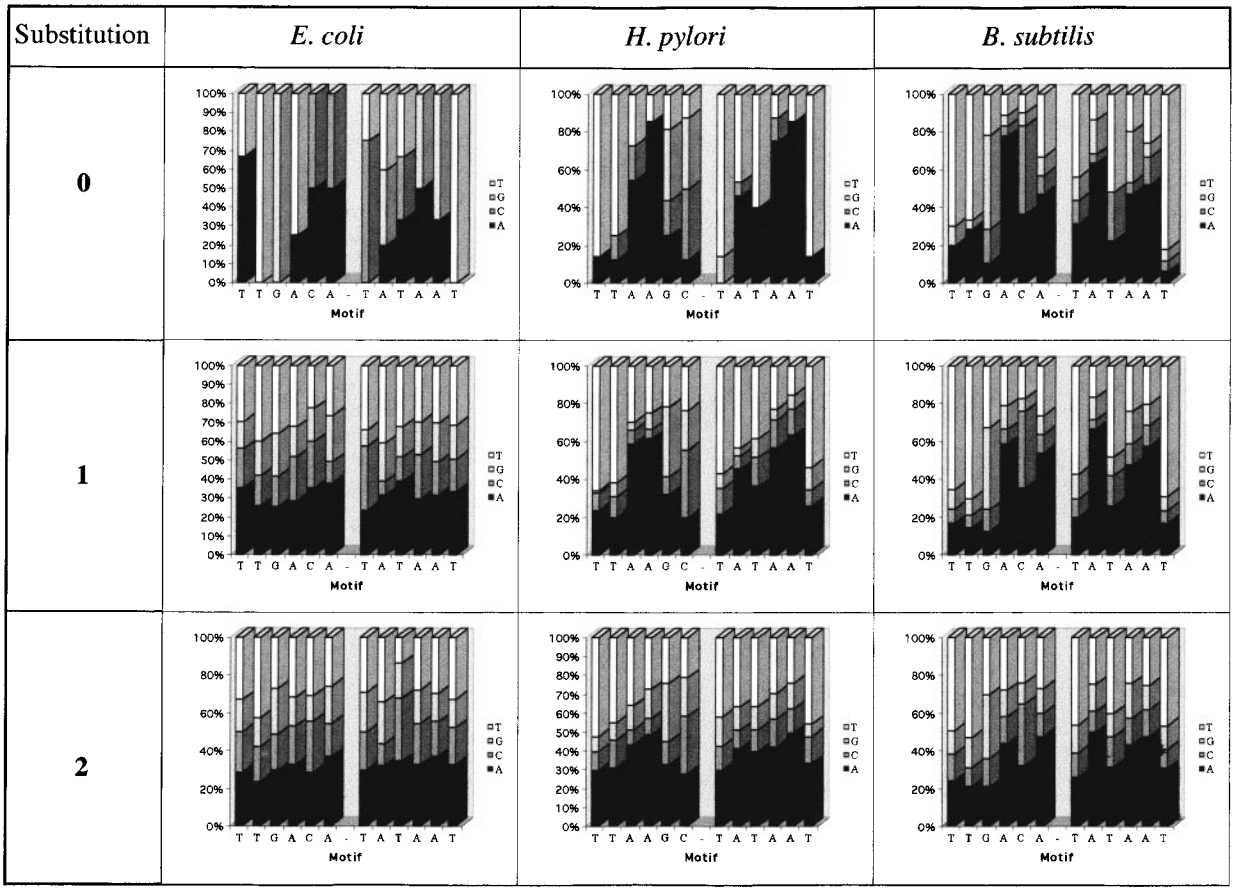

Figure 5. Frequency of occurrence in *H. pylori* of the motif TTAAGC×(19-23)TATAAT and in *E. coli* and *B. subtilis* of the motif TTGACA × 515-19)TATAAT with zero, one and two substitutions when each base in the motif is replaced, one at a time, by each one of the other possible bases in turn.

the pairs (TTTTAA, TATAAT) and (TTAAGC, TATAAT) (Figure 4), center around the values 12 and 21, respectively. These values are the same as those identified by plotting the χ^2 probability of occurrence of the most significant combined motifs extracted by algorithm 2 against the distances allowed between the two parts of the motifs (Figure 3).

The TTTTAA motif followed by TATAAT is frequent in the *H. pylori* genome. The distance between the two motifs peaks at 12. The smaller peak at 21-22 when up to two substitutions are allowed is probably due to an interference with TTAAGC×(19-23)TATAAT. A symmetrical phenomenon appears in the plot for TTAAGC×(19-23)TATAAT.

No motif, neither those found in *H. pylori*, nor the classical, TTGACA×(15-19)TATAAT, is significant in *E. coli* sets A or B, either in terms of frequency, or of distance between the two parts of the motifs. This is surprising, as the *E. coli* genome is approximately the same size as the *B. subtilis* genome and is believed to contain fewer promoter sequence families. However, using *E. coli* set C, the TTGACA×(15-19)TATAAT motif is identified as significant (χ^2 value of 10.41, just slightly above 10^{-3} in terms of probability). This motif appears with no error only once in sets A and B, and never in set C.

Checking for a possibly extended three-part motif for the σ^{80} promoter in H. pylori. It is possible for motifs TTTTAA×(10-14)TATAAT and TTAAGC×(19-23)TATAAT to occur together. We therefore tested for the statistical significance of the number of occurrences of TTAAGC (1-5)TTTTAA×(10-14)TATAAT and of the strictness of the distances separating any two elements of this putative three-part motif. Table 7 shows that the TTAAGC×(1-5)TTTTAA×(10-14)TATAAT motif is statistically significant for both aspects in *H. pylori*. The less frequent TTAAGC× (1-5)TTAAGC×(10-14)TATAAT motif is equally statistically significant. The central TTAAGC is specific to some promoters. Indeed, in the 23 S-5 S and 16 S RNA genes, the TTAAGC× (1-5)TTAAGC×(10-14)TATAAT motif appears with, respectively, zero and one substitution overall.

Table 7. Number and frequency of occurrence per sequence as well as statistical significance in sets A, B and C for *H. pylori*, *B. subtilis* and *E. coli* of the two possible extensions of the main two-part motifs identified by algorithm 2 in *H. pylori*

Subs.	Motif	*H. pylori* Set A	*H. pylori* Set B	*H. pylori* Set C	*E. coli* Set A	*E. coli* Set B	*E. coli* Set C	*B. subtilis* Set A	*B. subtilis* Set B	*B. subtilis* Set C
2	TTAAGC x(1-5) TTTTAA x(10-14) TATAAT	23 3.04 %	21 6.18 %	6 20%	4 0.15 %	2 0.17 %	0 0%	4 0.15 %	2 0.19 %	0 0%
3	TTAAGC x(1-5) TTTTAA x(10-14) TATAAT	101 13.36 %	77 22.65 %	17 56.67 %	34 1.28 %	16 1.34 %	0 0%	54 1.98 %	32 2.97 %	0 0%
4	TTAAGC x(1-5) TTTTAA x(10-14) TATAAT	345 45.63 % **85.14**	192 56.47 % **63.43**	23 76.67 % **8.48**	249 9.39 % **6.42**	127 10.64 % 1.35	23 5% 0.08	348 12.79 % 2.33	193 17.94 % **16.46**	0 0% 0.02
2	TTAAGC x(1-5) TTAAGC x(10-14) TATAAT	5 0.66 %	4 0.29 %	5 16.67 %	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
3	TTAAGC x(1-5) TTAAGC x(10-14) TATAAT	52 6.88 %	28 8.24 %	10 0.53 %	14 33.33 %	6 0.50 %	0 0%	17 0.62 %	9 0.84 %	0 0%
4	TTAAGC x(1-5) TTAAGC x(10-14) TATAAT	205 27.12 % **39.81**	106 31.18 % **20.65**	15 50% 3.36	163 6.15 % 0	98 8.21 % 0.24	1 2.70 % 0	185 6.80 % 0.51	107 9.94 % 1.11	0 0% 3.16

The motifs were searched with two, three and four substitutions. Bold, underscored values correspond to a χ^2 value whose probability of appearing by chance is less than 10^{-5}, bold values to a χ^2 value whose probability of appearing by chance is less than 5×10^{-2}. Statistical significance is indicated in the case of four substitutions only.

For TTAAGC×(1-5)TTTTAA×(10-14)TATAAT and TTAAGC×(1-5)TTAAGC×(10-14)TATAAT, the peaks in the distances between each pair of adjacent motifs is observed at 4 and 11, respectively (results not shown). These two distances, plus six (the length of TTTTAA or TTAAGC), total 21, which corresponds to the peak of the distances between TTAAGC and TATAAT.

List of the genes flanked in 5′ by the TTAAGC×(19-23)TATAAT motif. We then listed all *H. pylori* non-CDSs and all sequences in set C not present in set A that contain at least one occurrence of the TTAAGC×(19-23)TATAAT motif, with zero or one substitution. This revealed that seven such non-CDSs† have an occurrence of the motif with no error, and 49 with one substitution. The 56 sequences are 64 to 4549 nt long. Five of these 56 sequences, all longer than 1000 nt, were defined as non-CDSs by us, although they code for rRNA molecules or tRNA molecules.

A total of 17 of the 56 sequences are at the 5′ end of an operon coding for putative proteins. Ten of the other 39 sequences are upstream from genes coding for elements of the translational machinery, that is, of genes that are strongly expressed. Indeed, the two operons coding for both 23 S-5 S RNA genes possess the motif without error, and the two operons coding for both 16 S RNA genes present the motif with one substitution. Furthermore, the motif appears with zero or one substitution upstream from six ribosomal protein genes or operons. Fifteen of the 20 known genes or operons coding for other ribosomal proteins have the motif with two or three substitutions. The genes for both translational elongation factors (EFs) *EF-Tu* and *EF-Ts* are part of operons having upstream of their sequence the motif with up to one substitution. Two non-CDSs containing the motif with up to one substitution are also present in the *Cag* pathogenetic island (HP0536, HP0546).

† The total could be eight, but in one case, that of a non-CDS upstream from a phage/colicin/tellurite resistance cluster *terY* protein, the exact occurrence appears some 4200 bases before the start of transcription and is unlikely to be the gene's promoter.

Similarly, the motif is associated with genes coding for proteins that participate in important cell functions such as cell division, replication and transcription, and for proteins that are well expressed. These include the *fixNOQP* operon (HP0144) coding for the subunits of the cytochrome *c* oxidase that catalyse ATP formation induced by oxygen limitation (which is the case in *H. pylori*) (Preisig *et al*., 1993), the *hsdR/hsdM* system (HP1402, HP1403), which is a restriction enzyme system of type I, and various proteins with metabolic functions (especially nucleotide metabolism).

Five non-CDSs (HP0103, HP0325, HP0099, HP0231 and HP1559) are located upstream from genes coding for proteins involved in the flagellar biosynthesis and chemotaxis. These proteins are generally described as being transcribed from a promoter recognized by a σ^F factor (a σ factor that is believed to exist in *H. pylori* (Tomb *et al*., 1997)). The consensus sequence for the σ^{28} factor (homologous to the σ^F factor) was first described in *B. subtilis* (Gilman *et al*., 1981) as being CTAAA-N16-CCGATTA, and then in *E. coli* and *S. typhimurium* (Helmann & Chamberlin, 1987) as being TAAA-N15-GCCGATAA. We were not able to identify any such combined motif with our algorithm, possibly because it is too rare to have been detected with the higher quorums we used. However, the prefix of the GCCGATAA motif at the −10 site which is recognized by the σ^{28} protein of *E. coli* may be related to the suffix of the TTAAGC motif we suggest for *H. pylori*.

Finally, the σ^{80} gene in *H. pylori* possesses the motif TTAAGC×(*d*)TATAAT with one substitution and a spacing of 18 between the two parts.

Discussion

We tried to identify a consensus promoter motif in a given organism (in this case, *H. pylori*) by considering all non-coding sequences in its genome. This contrasts with the usual approach, which consists of using only a subset of sequences experimentally chosen for containing as promoter. We also tried to extract the consensus in as unbiased a way as possible; in particular, we did not make use of what is known about such motifs in other organisms. To this purpose, we used algorithms that can exhaustively extract from a set of sequences either single motifs or a pair of motifs separated by a distance belonging to an user-defined interval.

We are thus able to propose a combined motif, TTAAGC×(19-23)TATAAT for the σ^{80} promoter sequence in *H. pylori*. The box at −10 (TATAAT) is the same as that previously described for *E. coli* and *B. subtilis*. The box at −35 and the distance between the two boxes are different. We suggest also that a third motif, TTTTAA, or, sometimes, an approximate repetition of TTAAGC, may be located between these two at an optimal distance of 12 bp from the TATA-box. The frequency of the occurrences of both combined motifs and a simple statistical test suggest that the motifs are pertinent. The TTAAGC×(19-23)TATAAT and TTTTAA×(10-14)TATAAT motifs are indeed more frequent (31.35 % and 59.39 %, respectively, with two substitutions at most) in the non-CDS sequences (set A) than, for instance, the classical promoter consensus, TTGACA×(15-19)TATAAT, in either *E. coli* (9.31 %) or *B. subtilis* (23.30 %)†. Furthermore, Table 6 shows that these occurrences are statistically significant. The occurrences of the TTAAGC×(19-23)TATAAT and TTTTAA×(10-14)TATAAT motifs are not deemed statistically significant either in *E. coli* nor, in the case of the first motif, in *B. subtilis*. The second motif, TTTTAA×(10-14)TATAAT, seems significant in *B. subtilis* (less than in *H. pylori*) but this needs confirmation. On the other hand, the classical promoter consensus, TTGACA×(15-19)TATAAT, is statistically significant in *B. subtilis* but not in *H. pylori*, nor, more surprisingly, in the set of non-coding sequences extracted from *E. coli*. Indeed, running algorithm 2 on *E. coli* sets A and B and plotting the χ^2 of the most significant combined motif against the distances between the two parts of the motifs yielded a flat curve (result not shown). The classical motif is, however, well conserved, with a high enough level of statistical significance, just upstream from *E. coli* genes coding for ribosomal RNA molecules or ribosomal proteins. It is also reasonably well conserved, at very low quorums, in the experimental set recovered from Ozoline (1998) (see above). These observations suggest that such sequences present more variability in *E. coli* than in *B. subtilis*. In particular, the spacing between the two motifs appears less strict in *E. coli* than in *B. subtilis*.

Approximately one quarter of the genes coding for putative proteins in *H. pylori* and having upstream an occurrence with zero or one substitution of the TTAAGC×(19-23)TATAAT motif are elements of the translational machinery (ribosomal RNA molecules, ribosomal proteins and tRNA molecules). These elements are essential, and are abundant in the cells in exponential phase. This further supports the relevance of the motifs identified.

The sequence of the σ^{80} protein in *H. pylori* is the most distant when compared to that of the corresponding proteins in other bacteria (Solnick *et al*., 1997). A PILEUP analysis of 15 bacterial σ^{70} sequences (σ^{80} for *H. pylori*) showed that the 2.4 region in *H. pylori* (described as binding the −10

† The classical consensus TTGACA×(15-19)TATAAT that thus seems significant, at least statistically speaking, is less frequent in *B. subtilis* than out TTAAGC×(19-23)TATAAT motif in *H. pylori*. This may be because *B. subtilis* is believed to have more promoter families (possibly 18) than *H. pylori* (opossibly only three; (Tomb *et al*., 1997).

promoter region on the DNA (Lonetto *et al*., 1992) and comprising 22 amino acid residues) is identical with that of *E. coli* except for V → I and S → A substitutions. In contrast, the 4.2 region in *H. pylori* σ^{80} (described as binding the −35 promoter region and containing 29 amino acid residues) differs at nine positions from that in *E. coli* σ^{70}. The *H. pylori* σ^{80} protein is 30 amino acid residues longer at the N-terminal end of the protein than most σ^{70} proteins. These observations are in agreement with the fact that the −10 box that we describe for *H. pylori* is identical with that in *E. coli* whereas the −35 box is different.

Different hypotheses could be proposed concerning the presence of a TTTTAA motif 10 to 14 bp upstream from the TATA-box in *H. pylori*. This could be due to the composition of the genome, which is very A+T-rich. However, this is unlikely because the motif (A)AATT(T), which may be equivalently explained, was not identified as being statistically significant (result not shown). The structure may also represent an extended −35 motif. A similar extended motif has been suggested by Wosten *et al*., (1998b) for *C. jejuni* (see below). Moreover, two nucleotides located one base upstream from the −10 box have been described as an extended −10 motif and are known to bind the 2.5 region of the σ^{70} protein in *E. coli* (Barne *et al*., 1997). Even if the hypothesis of an extended −35 box were true, we do not know which part of the σ^{80} protein would bind to it. Finally, a third possibility is that the TTTTAA motif, and overall richness of A+T in the region between the two motifs, could bend the DNA to give a structure that would facilitate the binding of the RNA polymerase to the TTAAGC and TATAAT boxes which are approximately 21 nt apart.

Promoter sequences have been identified in *C. jejuni* (Wosten *et al*., 1998b), a bacterium closely related to *H. pylori*. Wosten *et al*. (1998b) characterized 11 promoters by cloning chromosomal DNA fragments upstream from a promoterless *lacZ* gene and transforming *C. jejuni* with this library. After having identified the transcriptional start of these 11 sequences, the authors aligned them with ten other previously characterized promoters to establish a consensus sequence. In most positions of the consensus, more than 50 % of the sequences contain the same residue. Wosten *et al*. thus found a TATAAT motif, like ours, for the TATA-box, and a TTAAGTxxTT motif at position −35, whose first five nucleotides are exactly the same as those of our motif, TTAAGC. The following dinucleotide, TT, could correspond to the beginning of the TTTTAA motif found in *H. pylori* (which is located one to five nucleotides downstream from the −35 motif). If the two T nucleotides in these positions form part of the −35 box in *H. pylori*, the distance between the two boxes (−35 and −10) would be 17 nt, in agreement with the distance between the −35 and −10 boxes in *E. coli* and *S. typhimurium.* Wosten also cloned and sequenced *C. jejuni*'s σ protein (Wosten *et al*., 1998a). The σ factor of *C. jejuni* encoded by the *rpoD* gene is 40 % identical with the corresponding protein in *E. coli,* and 66 % identical with the *H. pylori* σ^{80} (also encoded by the *rpoD* gene). The 2.4 region of the protein that binds to the −10 promoter sequence is conserved in *E. coli*, *C. jejuni* and *H. pylori*. In contrast, the 4.2 region that binds the −35 box is 94 % identical in *C. jejuni* and *H. pylori* but is just 56 % identical with that in *E. coli*. In particular, an isoleucine residue between the two arginine residues that bind the G and the C nucleotides in the classical −35 box is replaced in *C. jejuni* and *H. pylori* by a valine residue. The authors thus suggest that the −35 box in *H. pylori* may have a sequence similar to that of *C. jejuni*. This view is confirmed by our results.

Biological experimentation is required to confirm the validity of the results concerning *H. pylori* (equivalent experiments have been performed on *C. jejuni*, which in part confirm our findings). Several *H. pylori* genes are only weakly expressed from their own promoters in *E. coli* (Beier *et al*., 1998). This suggests that the whole *E. coli* RNA polymerase cannot efficiently recognize the *H. pylori* σ promoter sequence. It would therefore be interesting to verify whether the promoter corresponding to the consensus described here, TTAAGC×(19-23)TATAAT, is recognized by the *E. coli* σ protein. This could be achieved by cloning *H. pylori* promoters containing our motif in *E. coli* upstream from the *lacZ* gene. If, under these conditions, the promoter is not recognized, the experiment could be repeated with *E. coli* producing the *H. pylori* σ^{80} protein, with and without the core of the *H. pylori* RNA polymerase. If the promoter is activated, it would prove that the motif described here is recognized by the polymerase. Directed mutagenesis could also be used to change one or several bases in the motif, or alter the distance between the two boxes, and determine the effects on the promoter strength. Finally, it would be interesting to knock out the genes containing the TTAAGC×(19-23)TATAAT motif with one or two substitutions, especially those corresponding to putative proteins. As our motif is strongly conserved just upstream from these genes, the cell may require them to be well expressed, and they may thus code for proteins with important functions.

Although the approach used has proved to be useful, much remains to be done. In terms of algorithms, we are growing increasingly more sophisticated and efficient, even using combinatorial approaches as in this paper. We are, however, lagging behind in the statistical evaluation of the motifs found. This is specially true where errors (substitutions only or, more generally, susbtitutions, insertions and deletions) are allowed, and motifs may be composed of more than one part, with adjacent parts standing at specific distances. Typically in our case, identifying the motifs may take between a few seconds to a few minutes. The statistical evaluation using a data shuffling approach may take hours. This will happen when

the search for motifs is performed in a very flexible way (low quorums and high number of substitutions permitted) yielding numerous motifs (e.g. 2000-3000). Of course, this depends also on the number of shufflings one estimates are necessary to obtain reliable statistics.

Purely combinatorial approaches like our own (that may use statistics to evaluate the quality of the results obtained but *a posteriori* only) allow for a much more controlled and precisely defined analysis of the data. However, such approaches require extended practice in their use. In particular, setting the values for the various parameters (length of motifs, maximum number of substitutions allowed and quorum) is currently done by trial and error, or by running repeatedly the algorithm with different sets of values. The quorums used here correspond in general to the highest ones, or close to the highest at which significant motifs, or any motifs at all are found. With time, intuition develops on how best to set these values. Work must still be done to formalize and integrate such intuition directly into the algorithms. A first step in that direction has been made concerning the distance separating the two parts of a combined motif (Marsan & Sagot, 1999). Precisely defining and automating the grouping of motifs into families is also a not trivial task we need to address in the future.

Materials and Methods

The data

Test sets

The two test sets used for initially testing our algorithms consisted in sequences from *E. coli* and *B. subtilis* containing an experimentally determined transcript start or, sometimes, promoter. They were obtained from work by Ozoline *et al.*, 1998) and Helmann (1995), respectively. In both cases, the sequences are aligned on the start of transcription. Sequences for *E. coli* stop at that point, sequences for *B. subtilis* have 20 bp more downstream from the transcription start. The *E. coli* experimental set contains 441 sequences of average length 79 bp, for a total of 35,115 nucleotides. The *B. subtilis* set contains 131 sequences of average length 99 bp, for a total of 13,099 nucleotides. G+C content is 41 % for *E. coli*, and 32 % for *B. subtilis*.

H. pylori *data sets*

The initial data consisted of the set of non-coding sequences extracted from the whole genome of *H. pylori* that was sequenced and annotated at TIGR (Tomb *et al.*, 1997) and obtained from ftp://ftp.tigr.org/pub/data/h_pylori

We defined as non-coding, sequences corresponding to regions non-coding on both strands located upstream from genes (taking care of the direction of transcription of the described ORFs). The promoters for some genes may be located in sequences coding for other genes on either the same or the other strand. However, this situation is sufficiently rare that we chose to ignore these cases rather than run the risk of introducing in our data an unwarranted amount of noise and two possibly different compositional landscapes. We also chose to eliminate from the data all sequences of less than 40 bp long. We deemed that much space is necessary to accomodate (inside a single non-coding region) a promoter, as well as the transcription and translation start points. All sequences that contained a non-standard letter (usually an N) were also discarded, except where this letter appeared farther than 40 bases upstream from the start of translation. In such cases, we included the sequence, from the standard letter (A, C, G or T) to the right of the last N, to the start of translation (few sequences were discarded by this last operation).

The data obtained consisted of a set of 756 sequences (corresponding to set A) varying in length between 41 and 4549 bp (average 223), for a total of 168,709. These sequences are very A+T-rich (their content in G+C is less than 33 %). The overall G+C content of *H. pylori* is slightly less than 40 %. Some of the sequences may not contain any promoter (i.e. correspond to intergenic regions whose flanking genes are transcribed in convergent directions) while others may have two (i.e. correspond to intergenic regions whose flanking genes are transcribed in divergent directions). Regions between divergent genes appear twice in the set, once for each direction. Sets of sequences potentially less noisy than set A were also constructed.

We thus extracted from set A a subset, called set B, composed of the sequences located upstream from genes regulated by two divergent promoters, one on each strand. Set B comprises 340 sequences varying in length between 50 and 4549 bases (average 286), for a total of 97,360 nucleotides.

Finally, we established a third set, called set C. This third set consists of the non-coding regions upstream from genes coding for either a ribosomal RNA or a ribosomal protein, or from operons including ribosomal protein genes. Set C comprises 30 sequences varying in length between 44 and 749 bp (average 235), for a total of 7066 nt. It has a non-empty intersection with set B but is included neither in it nor in set A.

The main information concerning these data is summarized in Table 1.

The algorithmic approach

The algorithms

The two algorithms used to identify motifs for *H. pylori* promoters in a set of non-aligned sequences are based on the idea of inferring motifs by flexibly comparing words in the sequences to an external object, instead of between themselves (Sagot *et al.*, 1995, 1997; Sagot & Viari, 1996; Sagot, 1998). In our algorithmic papers, we used the term ''model'' to designate such objects. A model corresponds to what we have called so far a motif; it is a word over the DNA alphabet $\Sigma = \{A, C, G, T\}$ (Sagot *et al.*, 1995). A word u in a sequence is then said to be an e-occurrence (or, more simply, an occurrence) of a model m if the minimum number of errors (i.e. mutations corresponding to substitutions, deletions or insertions) necessary to transform u into m is no more than e, where e is a non-negative integer. Here, only substitutions were permitted. For instance, if m is the word ACG, then CCG, GCG, TCG, AAG, AGG, ATG, ACA, ACC, ACT are 1-occurrences (occurrences) of m. In the results given (Tables 1 to 7), models only are indicated, not their lists of occurrences.

Models are unknown before the algorithm is run, and are recursively constructed in an efficient way that allows us to consider only those constantly satisfying a certain constraint that is:

(i) A quorum q: meaning a valid model must have occurrences in at least q different sequences of the set (setting q at less than the total number N of sequences allows us to deal with noisy sets). A model may have more than one occurrence in a same sequence, only one is counted to verify whether the quorum is satisfied.

Final models must satisfy a further constraint, namely:

(ii) a length: a model length may vary between a minimum and a maximum value fixed by the user. The maximum value may be the largest for which the model still satisfies constraint (i).

A modification of the algorithms introduced by Sagot *et al*. (1995, 1997; Sagot & Viari 1996, 1997) allows two models to be constructed simultaneously, each one of which satisfies constraint (i), as well as constraint (ii) where final models are concerned, while the two together further satisfy the following:

(iii) let m and m' be the models; given any occurrence u of m in a sequence s of the set, there exists at least one occurrence u' of m' also in s, such that u comes before u' and the distance between the end position of u and the start position of u' is within an interval (*dmin*, *dmax*) where *dmin* and *dmax* are non-negative integers fixed by the user. An equivalent, symmetrical, property must be true also for all occurrences u' of m'.

The value of e determining whether a model has an occurrence in a sequence may be fixed at different levels for the two models. This allows us to ask if one model is better conserved than the other with which it is associated (as is, for instance, the case with the TATA-box in most bacterial promoters). An overall maximum number of errors may also be established. For instance, we may allow up to one substitution in either parts of a combined model, but only one overall. This means that if an occurrence of the first part presents one substitution, its companion (that is, an occurrence of the second part at a right distance) must be exact (and *vice versa*).

Algorithm 1 corresponds to the one given by Sagot *et al*., 1995, 1997; Sagot & Viari (1996). The models (motifs) it builds satisfy constraints (i) and (ii) but not (iii). Models (motifs) constructed by algorithm 2 satisfy all three constraints (Marsan & Sagot, 1999). Such models allow us to identify DNA binding sites recognized by the same protein (e.g. the RNA polymerase). These sites must, for steric reasons, stand at a defined distance apart.

A preliminary version of the algorithm is available upon request but only to non-profit research organizations. This represents a prototypal core algorithm, not a fully developed software. Its use requires careful understanding of the algorithm and various parameters asked as inputs. Suggestions from biologists concerning possible improvements to the algorithm (in particular as regards extension of its applicability to eukaryotes) are welcome.

Statistical evaluation of the motifs

There are two main types of approaches possible for the assessment of the statistical significance of the motifs found in a set of sequences that could have been adopted: one is based on some theoretical model of the sequences, the other corresponds to the data shuffling approach.

None of the methods of the first type is completely satisfactory for our purposes. Indeed, as we allow substitutions between motifs and their occurrences, we need methods that are able to either deal with such substitutions, or handle multiple exact motifs statistics. Those available (M. Régnier & W. Szpankowski, unpublished results; Reinert & Schbath, 1998; Tompa, 1999) appear too computationally intensive for our purposes. A further complicating factor comes from the fact that we are interested in assessing the statistical significance of the number of occurrences of a motif per sequence, i.e. of distinct sequences with at least one occurrence, and not of the total number of occurrences.

For these reasons, it seemed to us more appropriate to evaluate the pertinence of a motif by using the data-shuffling approach (Karlin *et al*., 1989).

The statistical significance of the models (motifs) found was thus evaluated by performing a χ^2 test (with one degree of freedom) on two contingency tables, one corresponding to what is observed, the other to what is expected under the null hypothesis (Press *et al*., 1993), and then determining the probability of getting the motifs observed, given the null hypothesis. A thousand random shufflings preserving both the mono and dinucleotide frequency distributions of the original set of sequences were performed to derive the values in the contingency table for the null hypothesis. Another type of statistic, based on a Z-score, was also used.

Acknowledgments

A.V. was supported by UPR 9073 of the CNRS, the French Ministry of Research (PRFMMIP to Philippe Régnier), the University of Paris VII, OraVax Inc. (Boston, MA, USA) and Pasteur Mérieux Connaught (Lyon, France). L.M. and M.-F.S. were partly supported by a CAPES/COFECUB project (of type II, number 272/99) between the universities of Marne-la-Vallée and Rouen in France and of São Paulo and Campinas in Brazil, as well as by the REMAG project with the INRIA, France. The authors thank Hilde de Reuse, Catherine Letondal, Laurent Bloch, Patrick Stragier and Alain Viari for their careful reading of the manuscript. They thank the referees for their very constructive comments that helped to improve the paper.

References

Bailey, T. L. & Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learn.* **21**, 51-80.

Baldi, P., Chauvin, Y. T. H. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA,* **91**, 1059-1063.

Barne, K. A., Bown, J. A., Busby, S. J. & Minchin, S. D. (1997). Region 2.5 of the *Escherichia coli* RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended-10' motif at promoters. *EMBO J.* **16**, 4034-4040.

Beier, D., Spohn, G., Rappuoli, R. & Scarlato, V. (1998). Functional analysis of the *Helicobacter pylori* principal sigma subunit of RNA polymerase reveals that

the spacer region is important for efficient transcription. *Mol. Microbiol.* **30**, 121-134.

Blaser, M. J. (1992). *Helicobacter pylori*: its role in disease. *Clin. Infect. Dis.* **15**, 386-391.

Cardon, L. R. & Stormo, G. D. (1992). Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.* **223**, 159-170.

Chen, Q. K., Hertz, G. Z. & Stormo, G. D. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.* **11**, 563-566.

Correa, P. (1995). *Helicobacter pylori* and gastric carcinogenesis. *Am. J. Surg. Pathol.* **19**, s37-43.

Crowley, E. M., Roeder, K. & Bina, M. (1997). A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.* **268**, 8-14.

Fraenkel, Y. M., Mandel, Y., Friedberg, D. & Margalit, H. (1995). Identification of common motifs in unaligned DNA sequences: application to *Escherichia coli Lpr* operon. *Comput. Appl. Biosci.* **11**, 379-387.

Galas, D. J., Eggert, M. & Waterman, M. S. (1985). Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. *J. Mol. Biol.* **186**, 117-128.

Gilman, M. Z., Wiggs, J. L. & Chamberlin, M. J. (1981). Nucleotide sequences of two *Bacillus subtilis* promoters used by *Bacillus subtilis* sigma-28 RNA polymerase. *Nucl. Acids Res.* **9**, 5991-6000.

Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* a-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucl. Acids Res.* **23**, 2351-2360.

Helmann, J. D. & Chamberlin, M. J. (1987). DNA sequence analysis suggests that expression of flagellar and chemotaxis genes in *Escherichia coli* and *Salmonella typhimurium* is controlled by an alternative sigma factor. *Proc. Natl Acad. Sci. USA,* **84**, 6422-6424.

Karlin, S., Ost, F. & Blaisdell, B. E. (1989). Patterns in DNA and amino acid sequences and their statistical significance. In *Mathematical Methods for DNA Sequences* (Waterman, M. S., ed.), pp. 133-158, CRC Press, Boca Raton, France.

Kinsella, N., Guerry, P., Cooney, J. & Trust, T. (1997). The *flge* gene of *Campilobacter coli* is under the control of the alternative sigma factor. *J. Bacteriol.* **179**, 4647-4653.

Krogh, A., M., B., Mian, I. S., Sjoelander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.

Lawrence, C. E. & Reilly, A. A. (1990). An expectation minimization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymzer sequences. *Proteins: Struct. Funct. Genet.* **7**, 41-51.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wooton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science,* **262**, 208-214.

Lonetto, M., Gribskov, M. & Gross, C. A. (1992). The sigma 70 family: sequence conservation and evolutionary relationships. *J. Bacteriol.* **174**, 3843-3849.

Marsan, L. & Sagot, M. F. (2000). Extracting structured motifs using a suffix tree - algorithms and application to promoter consensus identification. *RECOMB 2000, Tokyo, Japan,* **In the press**.

McColl, K. E. (1996). *Helicobacter pylori* infection and its role in human disease: an overview. *Pharm. World Sci.* **18**, 49-55.

Mengeritsky, G. & Smith, T. F. (1987). Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. *Comput. Appl. Biosci.* **3**, 223-227.

Ozoline, O. N., Deev, A. A. & Arkhipova, M. V. (1998). Non-canonical sequence elements in the promoter structure. cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. *Nucl. Acids Res.* **25**, 4703-4709.

Preisig, O., Anthamatten, D. & Hennecke, H. (1993). Genes for a microaerobically induced oxidase complex in *Bradyrhizobium japonicum* are essential for a nitrogen-fixing endosymbiosis. *Proc. Natl Acad. Sci. USA,* **90**, 3309-3313.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1993). Numerical Recipes. In *The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Queen, C., Wegman, M. N. & Korn, L. J. (1982). Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. *Nucl. Acids Res.* **10**, 449-456.

Record, M. T., Reznikoff, W. S., Craig, M. L., McQuade, K. L. & Schlax, P. J. (1996). *Escherichia coli* RNA polymerase ($E\sigma^{70}$), promoters, and the kinetics of the steps of transcription initiation. In *Escherichia coli and Salmonella* (Neidhardt, F. C., ed.), vol. 1, pp. 792-820, ASM Press, Washington DC.

Reinert, G. & Schbath, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comput. Biol.* **5**, 223-253.

Sagot, M. F. (1998). Spelling approximate repeated or common motifs using a suffix tree. In *LATIN'98: Lecture Notes in Computer Science* (Lucchesi, C. L. & Moura, A. V., eds), pp. 111-127, Springer-Verlag.

Sagot, M.-F. & Viari, A. (1996). A double combinatorial approach to discovering patterns in biological sequences. In *Combinatorial Pattern Matching* (Hirschberg, D. & Myers, E. W., eds) Lecture Notes in Computer Science, vol. 1075, pp. 186-208, Springer Verlag.

Sagot, M.-F., Escalier, V., Viari, A. & Soldano, H. (1995). Searching for repeated words in a text allowing for mismatches and gaps. In *Second South American Workshop on String Processing* (Baeza-Yates, R. & Manber, U., eds), pp. 87-100, University of Chile.

Sagot, M.-F., Viari, A. & Soldano, H. (1997). Multiple comparison: a peptide matching approach. *Theor. Comput. Sci.* **180**, 115-137.

Solnick, J. V., Hansen, L. M. & Syvanen, M. (1997). The major sigma factor (RpoD) from *Helicobacter pylori* and other Gram negative bacteria shows an enhanced rate of divergence. *J. Bacteriol.* **179**, 6196-6200.

Stormo, G. D. (1990a). Consensus patterns in DNA sequences. *Methods Enzymol.* **183**, 211-221.

Stormo, G. D. (1990b). Identifying regulatory sites from DNA sequence data. In *Biological Structure, Dynamics, Interactions and Expression. Proc 6th Conf. Biomol. Stereodynam* (Sarma, R. & Sarma, M. S., eds), pp. 103-111, Adenine Press.

Stormo, G. D. & Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA,* **86**, 1183-1187.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L. X., Kirkness, E. F. & Peterson, S., *et al*. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature,* **388**, 539-547.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Seventh International Symposium on Intelligent Systems for Molecular Biology*, pp. 262-271, AAAI Press, Heidelberg, Germany.

Ulyanov, A. V. & Stormo, G. D. (1995). Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions. *Nucl. Acids Res.* **23**, 1434-1440.

van Helden, J., André, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827-842.

Waterman, M. S., Arratia, R. & Galas, D. J. (1984). Pattern recognition in several sequences: consensus and alignment. *Bull. Math. Biol.* **46**, 515-527.

Worlfertstetter, F., Frech, K., Herrman, G. & Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequenes by a novel tuple search algorithm. *Comput. Appl. Biosci.* **12**, 71-80.

Wosten, M. M., Boeve, M., Gaastra, W. & van der Zeijst, B. A. (1998a). Cloning and characterization of the gene encoding the primary sigma-factor of *Campylobacter jejuni*. *FEMS Microbiol. Letters,* **162**, 97-103.

Wosten, M. M., Boeve, M., Koot, M. G., van Nuene, A. C. & van der Zeijst, B. A. (1998b). Identification of *Campilobacter jejuni* promoter sequences. *J. Bacteriol.* **180**, 594-599.

Edited by M. Yaniv

(Received 3 September 1999; received in revised form 12 January 2000; accepted 4 February 2000)