Genomics

Differential genome analysis applied to the species-specific features of *Helicobacter pylori*

Martijn Huynen[a,b], Thomas Dandekar[a,b,*], Peer Bork[a,b]

[a] *EMBL, Meyerhofstr. 1, 69012 Heidelberg, Germany*
[b] *Max-Delbrück-Center, Berlin, Germany*

Received 19 January 1998; revised version received 25 February 1998

Abstract We introduce a simple and rapid strategy to identify genes that are responsible for species-specific phenotypes. The genome of a species that has a specific phenotype is compared with at least one, closely related, species that lacks this phenotype. Homologous genes that are shared among the species compared are identified and discarded from the list of candidates for species-specific genes. The process is automated and rapidly yields a small subset of the genome that likely contains genes responsible for the species-specific features. Functions are assigned to the genes, and dubious annotations are filtered out. Information is extracted not only from the presence of genes, but also from their absence with respect to known phenotypes. We have applied the technique to identify a set of species-specific genes in *Helicobacter pylori* by comparing it with its closest relatives for which complete genome sequences are available, *Haemophilus influenzae* and *Escherichia coli*. Of the genes of this set for which functional features can be obtained, a large fraction (63%, 123 proteins) is (potentially) involved in *H. pylori*'s interaction with its host. We hypothesize that a family of outer membrane proteins is critical for the ability of *H. pylori* to colonize host cells in highly acidic environments.

Key words: Comparative genome analysis; Species-specific phenotype; Sequence analysis; Outer membrane protein; *Helicobacter pylori*

1. Introduction

The recent publication of the entire genomic sequence of *Helicobacter pylori* [1] provides a wealth of sequence data. Suitable analysis of the genome allows the tracing of many metabolic and cellular processes of this pathogenic organism and will certainly open new roads in the treatment of peptic ulcers. Here we introduce a computational approach to identify genes that are specific to a species and hence are likely to be responsible for the specific features of its phenotype. Because of their species specificity, such genes are good drug target candidates. By comparing *H. pylori* to the pathogenic *Haemophilus influenzae* and to the rather benign *Escherichia coli* K12 this approach combined with comparative sequence analysis reveals genes that are responsible for (1) pathogen-specific features and (2) *H. pylori*-specific features such as acid tolerance which is essential for survival in the gastric environment.

*Corresponding author: any of the above. Fax: (49) (6221) 387-517.
E-mail: dandekar@embl-heidelberg.de

2. Materials and methods

All the protein coding genes of the three genomes were compared with each other (pairwise comparisons; each gene against each gene) to detect homologous relationships, using the rigorous Smith-Waterman algorithm [2]. The algorithm was run on a Biocellerator machine, which has parallel hardware and is specifically built to run dynamic programming algorithms (for more information about the Biocellerator see: http://www.cgen.com). Significant protein sequence similarities ('hits') were selected using an E value cut-off of 0.01. E values for protein sequence similarities are calculated as in Pearson's FASTA package, they are based on the expected number of pairwise sequence similarities of a certain level, given the sizes of the genomes compared (see [3] and references therein for a more detailed discussion). E values have been shown to be an accurate indication for the ratio of false positives to true positives of homologous relationships [4].

Homologous relationships between proteins of *H. pylori* and of *E. coli* [5] and *H. influenzae* [6] were noted and subsequently divided into orthologous and non-orthologous relationships. Homologous proteins are orthologs of each other when their sequence divergence reflects a speciation rather than a gene duplication event [7]. The division between orthologs and non-orthologs is made because orthologs are likely to perform the same function in the various species, having diverged relatively recently in evolution. To detect orthology, we use levels of sequence identity. Proteins are called orthologs of each other if (1) they have the highest level of pairwise similarity, compared to the identities of either gene to all the other genes in the other genome, (2) the similarity extends over 60% of the sequence of at least one protein (to exclude single domain hits in multi-domain proteins, but to include the possibility of gene fusion/splitting), and (3) the similarity is significant ($E < 0.01$). The *H. pylori* proteins that showed no detectable homology to *E. coli* or *H. influenzae* proteins were subsequently analyzed for homology to the proteins in EMBL database release 43 [8] and SwissProt database release 34 [9] using BLAST [10] and were further analyzed using profile search techniques described in [11]. *H. pylori*, *E. coli* and *H. influenzae* protein coding sequences were taken from the NCBI server at ncbi.nlm.nih.gov.

3. Results and discussion

In this paper we subtract various fractions from the *H. pylori* genome, and subsequently interpret their genes. A more general approach to the 'triple genome' comparison is shown in Fig. 1. Here the genes from the three genomes are divided into fractions that (1) have orthologs in all three species, (2) have orthologs in two of the genomes, but no homolog in the third genome, or (3) are unique to one of the three genomes (have no homolog in either of the other two genomes). Some of the functions of the genes that are unique to two of the genomes are highlighted in the figure. The three fractions in the figure do not add up to the size of the genome, since genes that only have non-orthologous homologs in other genomes were left out of the analysis.

After comparing the *H. pylori* genome with *E. coli* and *H. influenzae*, two other Gram-negative bacteria, we extracted

PII S0014-5793(98)00276-2

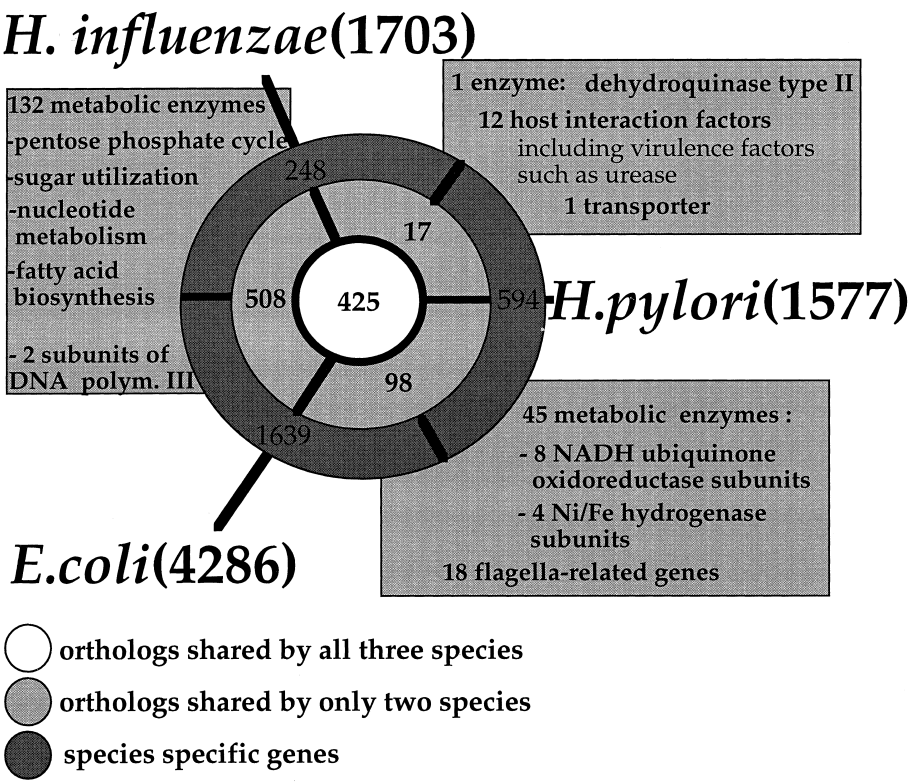

Fig. 1. Graphical representation of a triple genome comparison between *E. coli*, *H. influenzae* and *H. pylori*. The center of the circle gives the number of genes that have orthologs in all three species. The second ring gives the number of genes that have orthologs in two of the three species, but that do not have a homolog in the third. The outer ring gives the number of genes that are only present in one of the three species (have no homolog in the other two species).

and analyzed proteins that (A) have a homolog in the rather benign *E. coli* K12 laboratory strain, (B) have a homolog in the pathogenic *H. influenzae* but not in *E. coli* K12, or (C) have no homolog in either of the two.

(A) Of the set of 952 sequences subtracted as homologs to *E. coli*, 665 can be considered orthologs to genomic *E. coli* sequences, based on relative similarity (see above, [12]). Out of this set, 608 have a putative annotation. A relatively small fraction of these (69 proteins, 11%) can be considered 'host interaction factors' [12]. We define as host interaction factors as (1) proteins that are secreted or are involved in the last steps of the biosynthesis of such proteins, (2) receptors or other transmembrane proteins with extracellular regions, (3) known virulence factors (see also Table 1). Note, however, that not all genes involved in phenotypic characteristics of *H. pylori* are in the 'non-*E. coli*' set. For example, *H. pylori* shares with *E. coli*, but not with *H. influenzae*, a set of 18 genes for its flagella.

(B) The next fraction, 32 proteins that have homologs in the pathogenic *H. influenzae* but not in *E. coli*, consists mainly of host interaction factors. Of the 22 proteins of this set for which functional information was available or could be obtained by sequence analysis, 17 (77%) could be classified as host interaction factors (for the complete list see: http://www.bork.embl-heidelberg.de/Genome/Hpylori.html). Among these are five proteins of the urease operon. The urease operon has been implicated in pathogenicity, and in increasing the local pH of the environment to one more favorable for growth (reviewed in [13]). Specifically it is required for growth in the highly acidic environment of *H. pylori* [14], although it is not necessarily the only adaptation of *H. pylori* to the high acidity. Of the set of 32 proteins that have homologs in the pathogenic *H. influenzae* but not in *E. coli*, 17 can be considered orthologs.

This 'non-*E. coli*' set was checked for possible non-orthologous gene displacements [15] to avoid misprediction of the absence or presence of function. A non-orthologous gene displacement is exemplified by the only metabolic enzyme that

Table 1
Identification of species-specific genes in the *H. pylori* genome

H. pylori proteins	1577
Subtract *E. coli* hits (952)	625
Subtract *H. influenzae* hits (32)	594
Subtract sequences with unknown functions and without database hits (336)	258
Subtract hits only to proteins with unknown functions (62)	196
Subtract species-specific host interaction factors (123)	
Remaining set of species-specific genes	73

Table 2
The 73 genes that encode *H. pylori*-specific phenotypic features and that were not classified as host interaction factors

HP0037	Z	NADH-ubiquinone oxidoreductase subunit
HP0091	C	type II restriction enzyme (*hsdR*)
HP0111	R	similar to heat-inducible transcription repressor HrcA
HP0134	Z	3-deoxy-D-arabino-heptulosonate-7-phosphatase (*dsh1*)
HP0146	Z	cbb3-type cytochrome *c* oxidase subunit Q CcoQ
HP0147	Z	cytochrome *c* oxidase (*fixP*)
HP0168	Z	protein disulfide isomerase
HP0193	Z	fumarate reductase, cytochrome *b* subunit (*frdC*)
HP0200	D	ribosomal protein L32 (*rpl32*)
HP0219	Q	similar to an accessory factor for ABC transporters
HP0261	K	ribosomal protein-like, RNA binding
HP0275	C	ATP-dependent nuclease (*addB*)
HP0339	Z	lysozyme
HP0377	Z	thiol:disulfide interchange protein (*dsbC*)
HP0383	IK	signal-transducing protein, histidine kinase
HP0425	CI	ss DNA-specific exonuclease (*recJ*)
HP0431	Z	protein phosphatase 2C homolog (*ptc1*)
HP0432	Z	protein kinase C-like protein
HP0435	D	similar to *recJ*
HP0437	D	IS605 transposase (*tnpA*)
HP0454	Z	adenylate cyclase/kinase, substrate unknown
HP0470	Z	oligoendopeptidase F (*pepF*) (*Borrelia*)
HP0481	CD	adenine-specific DNA methyltransferase
HP0483	CD	cytosine-specific DNA methyltransferase
HP0589	Z	ferredoxin oxidoreductase, α subunit
HP0591	Z	ferredoxin oxidoreductase, γ subunit
HP0602	C	endonuclease III
HP0656	Z	Tim barrel, probably enzyme
HP0658	K	PET112-like protein, translation in *Borrelia*
HP0669	C	putative adenine-specific DNA methylase
HP0695	Z	hydantoin utilization protein, amino acid synthesis (*huyA*)
HP0696	Z	*N*-methylhydantoinase, amino acid synthesis (*serC*)
HP0736	Z	phosphoserine aminotransferase, amino acid synthesis
HP0741	K	protein of the HIT family
HP0800	K	molybdopterin converting factor (*moaE*)
HP0827	D	ss DNA binding protein 12RNP2 protein
HP0830	Z	amidase, *Borrelia*
HP0849	C	anti-codon nuclease masking agent
HP0988	D	IS605-like transposase (*tnpA*)
HP0998	D	IS605-like transposase (*tnpA*)
HP1000	D	PARA protein, plasmid partitioning
HP1004	DI	transposon gene (*mocA*)
HP1006	J	conjugal transfer protein (*traG*)
HP1022	D	DNA polymerase I (*polA*)
HP1077	QM	nickel transport protein (*nixA*)
HP1081	K	putative neuraminyllactose binding, flagella
HP1096	D	IS605-like transposase (*tnpA*)
HP1135	Z	ATP synthase F1, subunit δ (*atpH*)
HP1150	K	MLCB250.30, potential nucleic acid binding
HP1186	Z	carbonic anhydrase, pH regulation
HP1208	CD	ulcer-associated adenine-specific methyltransferase
HP1209	CD	ulcer-associated gene restriction enzyme (*iceA*)
HP1224	Z	uroporphyrinogen III cosynthase (*hemD*)
HP1227	Z	cytochrome c_{553}
HP1236	Z	ATP-dependent clp protease A, *Borrelia*
HP1238	Z	aliphatic amidase (*aimE*), can use acetamide as carbon source
HP1321	K	conserved hypothetical ATP-binding protein
HP1354	C	putative adenine-specific DNA methylase
HP1361	J	competence locus E (*comE3*)
HP1366	C	type IIS restriction enzyme R protein
HP1399	Z	arginase (*rocF*), arginine catabolism, urea cycle
HP1430	K	conserved hypothetical ATP-binding protein
HP1471	C	type IIS restriction enzyme R protein
HP1472	C	type IIS restriction enzyme M protein
HP1481	Z	similar to β-alanine synthetase
HP1496	K	general stress protein *Borrelia*
HP1507	K	conserved hypothetical ATP-binding protein
HP1521	C	type III restriction enzyme R protein
HP1530	Z	purine nucleoside phosphorylase (*purB*) *Borrelia*
HP1533	Z	probable 2,3-dihydrodipicolinate N-C6-lyase
HP1535	D	IS605 transposase-like (*tnpA*)
HP1539	Z	ubiquinol cytochrome *c* oxidoreductase, cytochrome *b* subunit
HP1540	Z	ubiquinol cytochrome *c* oxidoreductase, 2FE-2S subunit

The first column gives the *H. pylori* gene number, next the functional category is given. There are 31 enzymes (Z), 15 restriction enzymes (C), one regulatory protein (R), 11 proteins involved in interaction with DNA/RNA such as transposase (D), two transport proteins (Q), one internal hit to *H. pylori* only, which is thought to be a histidine kinase or a signal transducer (I), two proteins with conjugative function (J), and 10 with other functions (K). M denotes a metallo-protein. Relations to other categories are given by a second character. The third column gives a short description of the protein function. Proteins in this set that have a homolog in the pathogen *Borrelia burgdorferi* are denoted '*Borrelia*'.

H. pylori shares with *H. influenzae* and that has no homolog in *E. coli*, type II dehydroquinase (HP1038). The function of this enzyme is encoded in *E. coli* by an unrelated type I dehydroquinase [16].

(C) The automatic subtraction of the genes shared with *E. coli* and *H. influenzae* rapidly yields a subset of the complete *H. pylori* genome which is likely to be responsible for its species-specific features. A more detailed analysis of this smaller set is now possible. Appropriate functional classification is the second step to clearly identify the species-specific genes. Of the remaining 594 *H. pylori* genes with no homology either to *E. coli* or to *H. influenzae*, 258 are homologous to other known proteins. For 196 of these some functional classification was possible and they were studied in more detail.

In this set of 196 proteins, a large fraction has functions (see next paragraph) as host interaction factors (123 protein genes, 63%, for the complete list see the web page) (Table 1). Some of the other 73 genes hint at additional cellular features that are related to the specific environment in which *H. pylori* lives. They include 31 metabolic enzymes, 15 restriction enzymes as well as eight transposases and conjugation factors (Table 2). Some of the enzymes in this set may be specific adaptations of *H. pylori* to its environment. A carboanhydrase (HP1186) and an amidase (HP0830) can play a role in increasing the local pH, while an arginase (HP1399) produces urea which in turn can be used by the urease to increase the local pH.

Another example is a pyruvate ferredoxin oxidoreductase which catalyzes the transformation of pyruvate into acetyl-CoA. None of the four subunits (HP1108, HP1109, HP1110 and HP1111) of the enzyme has an ortholog in *E. coli* or *H. influenzae* whereas only two have (distant) homologs. Outside of *H. pylori*, orthologs of pyruvate ferredoxin oxidoreductase have only been observed in the Archaea and in *Thermotoga maritima* [17] (M.A. Huynen and P. Bork, unpublished results). The enzymes that catalyze the transformation of pyruvate into acetyl-CoA in *H. influenzae* and *E. coli* are either aerobic (*ace*AF) or anaerobic (*pfl*). The presence of the archaeal pyruvate ferredoxin oxidoreductase in *H. pylori* might

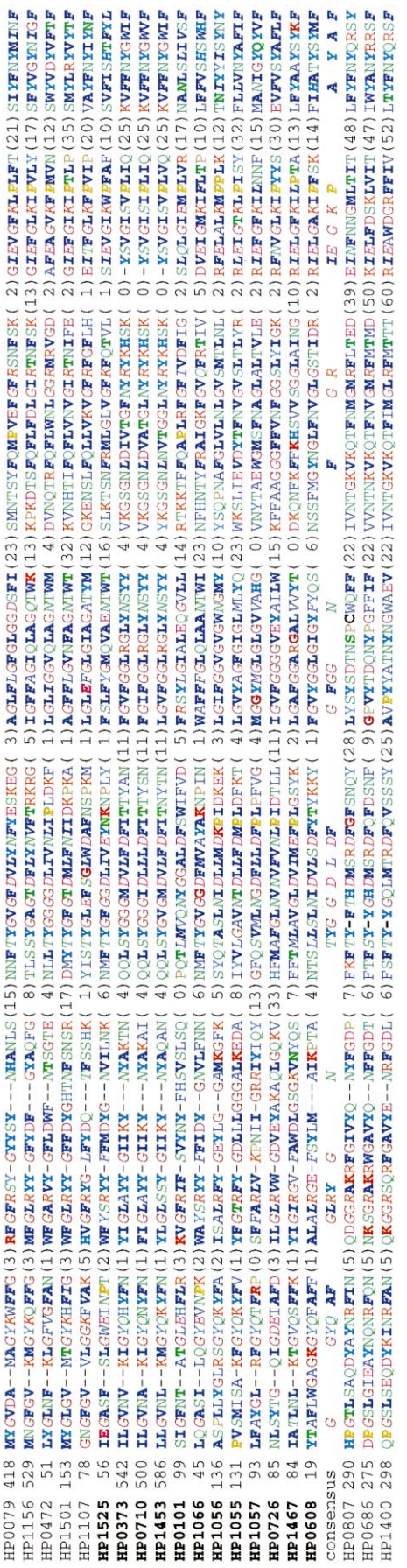

Fig. 2. Alignment of conserved regions of 12 new OMP members (bold) to selected known family members [1] and with three *H. pylori* iron(III) dicitrate transporters (bottom). Residues conserved in at least eight sequences (italics) are shown in the consensus line (a, aromatic residues; W,Y,F). Hydrophobic positions are given in bold. The family was expanded using iterative motif and profile searches (for details see [11]). One of the iron(III) dicitrate transporters, HP0807, scored considerably above the background of other database proteins [29].

hence be related to its micro-aerobic environment [1,17]. Interestingly, the enzyme that in *H. influenzae* and *E. coli* catalyzes the transformation of phosphoenolpyruvate into pyruvate in glycolysis, pyruvate kinase (*pykA* and *pykF* in *E. coli, pykA* in *H. influenzae*), is absent in *H. pylori*. Pyruvate kinase activity is in general most sensitive to changes in pH, K^+ and Mg^{2+} concentrations compared to other glycolytic enzymes [18]. This may be one of the reasons why this enzyme is missing in *H. pylori*. An enzyme which may have taken over the role of pyruvate kinase in *H. pylori* is phosphoenolpyruvate synthase [1].

The 123 potential host interaction factors include 23 out of the 27 genes in the cag pathogenicity island which has specifically been implied in *H. pylori* pathogenesis but is not necessary for survival [19]. We hypothesize that a large fraction of the remaining genes should be involved in adaptation to the most dramatic difference in the environment of *H. pylori* relative to that of both *E. coli* and *H. influenzae*: the high acidity of the gastric mucosa (pH = 2). The survival in the gastric environment seems to be the most complex aspect of the phenotype specific to *H. pylori* and would require a specific machinery. Thus, we looked for large sets of proteins that might perform related functions. The largest family within the remaining genes are outer membrane proteins (OMPs) [1]. We could enlarge the family by systematic sequence comparison from 32 [1] to at least 44 members and found that they are similar to iron(III) dicitrate receptors (Fig. 2), members of yet another family of porins [20]. This suggests that the family is involved in iron uptake. Specific iron uptake proteins have been shown to be located in the outer membrane of *H. pylori* [21]. Porins have been shown to be ligand-specific [22], so different porins should exist for different iron-associated ligands. For example, a 70 kDa protein has been identified as a transporter highly specific for human lactoferrin [23]. Functions that exclude iron uptake have been assigned for 41 of the 50 *H. pylori* proteins with a calculated molecular weight between 65 and 75 kDa (data not shown); the remaining nine all belong to this large OMP family.

Iron is not very soluble at the high acidity in the stomach [24], which might explain the presence of a specific, large gene family involved in iron uptake in *H. pylori*. Clinical observations that link iron to *H. pylori* pathogenicity and survival in the stomach include (i) correlation of lactoferrin levels in gastric mucosa with *H. pylori* infection and inflammation [25], (ii) involvement of the *H. pylori* outer membrane in pathogenic adhesion and iron-mediated attack to intestinal cells under acidic conditions [26], and (iii) reversal of iron deficiency anemia after eradication of *H. pylori* in superficial gastritis [27].

Although there is no direct evidence for the involvement of the OMP family in iron uptake, two of its members (HP1243, HP0896) have recently been shown experimentally to be adhesins that are important in determining host specificity, and were recommended as important vaccine targets [28]. The expanded number of putative iron uptake OMPs, relative to *E. coli* and *H. influenzae*, without an increase in inner membrane transporters (data not shown) might even indicate that the accumulated ion concentration in the periplasm or at the cell surface also contributes to the survival of *H. pylori* in the acidic environment.

Acknowledgements: This research was supported by the DFG (Bo 1099/3-1) and by BMBF (0300/401/1962D).

References

[1] Tomb, J.-F. et al. (1997) Nature 388, 539–547.
[2] Smith, T. and Waterman, M.S. (1981) J. Mol. Biol. 147, 195–197.
[3] Pearson, W. (1996) Methods Enzymol. 266, 227–258.
[4] Brenner, S.E., Hubbard, T., Murzin, A. and Chotia, C. (1995) Nature 378, 140.
[5] Blattner, F.E. and Bloch, G.P. et al. (1997) Science 277, 1453–1462.
[6] Fleishmann, R., Adams, M. and White, O. et al. (1995) Science 269, 496–512.
[7] Fitch, W.M. (1970) Syst. Zool. 19, 99–110.
[8] Stoesser, G., Moseley, M.A., Sleep, J., McGowran, M., Garcia-Pastor, M. and Sterk, P. (1998) Nucleic Acids Res. 26, 8–15.
[9] Bairoch, A. and Apweiler, R. (1998) Nucleic Acids Res. 26, 38–42.
[10] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[11] Bork, P. and Gibson, T.J. (1996) Methods Enzymol. 266, 162–184.
[12] Huynen, M., Diaz, J. and Bork, P. (1997) Trends Genet. 13, 389–390.
[13] Collins, C.M. and D'Orazio, S.E.F. (1993) Mol. Microbiol. 9, 907–913.
[14] Labigne, A. and de Reuse, H. (1996) Infect. Agents Dis. 5, 191–202.
[15] Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Trends Genet. 12, 334–336.
[16] Dijkhusen, L. (1996) in: Evolution of Microbial Life (Roberts et al., Eds.), pp. 243–266, Cambridge University Press, Cambridge.
[17] Hughes, N.J., Chalk, P.A., Clayton, C.L. and Kelly, D.J. (1995) J. Bacteriol. 177, 3953–3959.
[18] Podesta, F.E. and Plaxton, W.C. (1992) Biochim. Biophys. Acta 1160, 213–220.
[19] Covacci, A., Falkow, S., Berg, D.E. and Rappuoli, R. (1997) Trends Microbiol. 5, 205–208.
[20] Schulz, G. (1996) Curr. Opin. Struct. Biol. 6, 485–490.
[21] Worst, D.J., Sparrius, M., Kuipers, E.J., Kunster, J.G. and de Graaff, J. (1996) FEMS Microbiol. Lett. 144, 29–32.
[22] Jiang, X., Payne, M.A., Cao, Z., Foster, S.B., Feix, J.B., Newton, S.M. and Klebba, P.E. (1997) Science 276, 1261–1264.
[23] Dhaenens, L., Szczebara, F. and Husson, M.O. (1997) Infect. Immun. 65, 514–518.
[24] Cremonesi, P., Strada, D., Galimberti, G. and Sportoletti, G. (1984) Arzneim. Forsch. 34, 948–952.
[25] Nakao, K., Imoto, I., Ikemura, N., Shibata, T., Takaji, S., Taguchi, Y., Misaki, M., Yamauchi, K. and Yamazaki, N. (1997) Am. J. Gastroenterol. 92, 1005–1011.
[26] Corthesy-Theulaz, I., Porta, N., Pringault, E., Racine, L., Bogdanova, A., Kraehenbuhl, J.P., Blum, A.L. and Michetti, P. (1996) Infect. Immun. 64, 3827–3832.
[27] Marignani, M., Angeletti, S., Bordi, C., Malagnino, F., Mancino, L., Delle Fave, G. and Annibale, B. (1997) Scand. J. Gastroenterol. 32, 617–622.
[28] Ilver, D., Arnqvist, A., Ogren, J., Frick, I.M., Kersulyte, D., Incecik, E.T., Berg, D.E., Covacci, A., Engstrand, L. and Boren, T. (1998) Science 279, 373–377.
[29] Birney, E., Thompson, J.D. and Gibson, T.J. (1996) Nucleic Acids Res. 24, 2730–2739.