Molecular Microbiology (1999) **32**(4), 881–891

MicroCorrespondence

Penicillin tolerance in *Streptococcus pneumoniae*, autolysis and the Psa ATP-binding cassette (ABC) manganese permease

Sir,

Recently, Novak *et al*. (1998, *Mol Microbiol* **29:** 1285–1296) reported their investigation on the phenomenon of penicillin tolerance in *Streptococcus pneumoniae*. A library of mutants in pneumococcal surface proteins was screened for the ability to survive in the presence of 10 × the minimum inhibitory concentration of antibiotic. A mutant harbouring an insertion in the known gene *psaA* was isolated among 10 candidate tolerance mutants. Inactivation of *psaA* was previously shown to result in reduced virulence of *S. pneumoniae* (as judged by intranasal or intraperitoneal challenge of mice) and in reduced adherence to A549 cells (type II pneumocytes), leading to the suggestion that PsaA was an adhesin (Berry and Paton, 1996, *Infect Immun* **64:** 5255–5262). This gene is part of the *psa* locus (Fig. 1) that encodes an ATP-binding cassette (ABC) permease belonging to cluster 9, a family of ABC metal permeases (Dintilhac *et al*., 1997, *Mol Microbiol* **25:** 727–740).

Novak *et al*. (1998, *Mol Microbiol* **29:** 1285–1296) reported that *psa* mutants displayed pleiotropic phenotypes: (i) reduced sensitivity to the lytic and killing effects of penicillin; (ii) growth in chains of 40–50 (*psaC*) to 200–300 (*psaD*) cells; (iii) autolysis defect and loss of sensitivity to low concentrations of deoxycholate (DOC), a species characteristic trait; (iv) absence of LytA, the major autolytic amidase; (v) almost complete loss of choline-binding proteins (ChBPs) (*psaC* and *psaD*) and absence of CbpA; (vi) loss of transformability (except *psaA*); and (vii) manganese (Mn) requirement for growth in a chemically defined medium.

Because penicillin tolerance was first associated with an autolysis defect (Tomasz *et al*., 1970, *Nature* **227:** 138–140), the absence of LytA (phenotype iv) could itself explain phenotypes i and iii. Dysregulation of *lytA* could not be investigated because, according to Novak *et al*. (1998, *Mol Microbiol* **29:** 1285–1296), the difficulty in lysing *psa* mutant cells prohibited Northern analysis, although lysates of the *psa* mutants could be obtained for immunoblot analysis of LytA and of RecA and for Southern confirmation of the *psa* mutations. Nevertheless, because expression of the *lytA* gene has been shown to be driven by three different promoters, including Pb which is the recA basal promoter (Mortier-Barrière *et al*., 1998, *Mol Microbiol* **27:** 159–170), and because wild-type levels of RecA were detected in the *psa* mutants (Novak *et al*., 1998, *Mol Microbiol* **29:** 1285–1296), it seems difficult to account for the complete absence of LytA on the basis of altered expression.

On the other hand, phenotypes i–iv are reminiscent of alterations observed after the replacement of choline (Ch) by ethanolamine (EA) in the cell wall of pneumococcus (Tomasz, 1968, *Proc Natl Acad Sci USA* **59:** 86–93). Similar phenotypes were also displayed by Ch-independent mutants of *S. pneumoniae* (Severin *et al*., 1997, *Microb Drug Res* **3:** 391–400; Yother *et al*., 1998, *J Bacteriol* **180:** 2093–2101). *S. pneumoniae* has a nutritional requirement for Ch that is incorporated by covalent bonds into the cell wall teichoic acids (TA) and in the membrane-bound lipoteichoic acid (LTA). Ch residues bound to TA (ChTA) were shown to be absolutely required for LytA activity (Holtje and Tomasz, 1975; *J Biol Chem* **250:** 6072–6076). The action of LytA has long been thought to be restricted to pneumococcal cell walls because of this requirement. However, recent reports suggest that ChTA is required

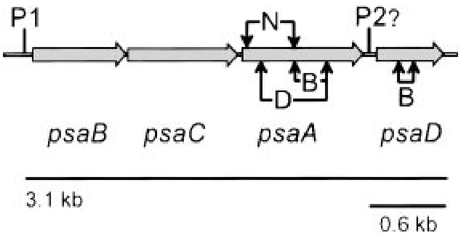


Fig. 1. Organization of the *psa* locus. ORFs are indicated by shaded arrows; *psaB*, *psaC*, *psaA* and *psaD* encode an ATP-binding cassette (ABC) protein, a hydrophobic membrane protein, a Mn-binding lipoprotein and a putative thiol peroxidase respectively. Arrows above and below the map define the limits of IDM-targeting fragments used for disruption of the *psaA* or the *psaD* (ORF3) genes, respectively, by: N, Novak *et al*. (1998, *Mol Microbiol* 29: 1285–1296); B, Berry and Paton (1996, *Infect Immun* 64: 5255–5262); D, Dintilhac *et al*. (1997, *Mol Microbiol* 25: 727–740). The 3.1 kb and the less abundant 0.6 kb transcripts detected with a *psaD* probe (Novak *et al*. 1998, *Mol Microbiol* 29: 1285–1296) are also indicated. It has been suggested that transcription of the *psa* locus is initiated from P1, a putative promoter located upstream of *psaB*. A second weaker promoter, P2 located immediately upstream of *psaD*, was postulated by Novak *et al*. (1998, *Mol Microbiol* 29: 1285–1296) to account for the shortest transcript. However, the latter could as well result from the processing of the 3.1 kb transcript. Taking into account the respective amount of each transcript and depending on the existence of P2, the expression of *psaD* could be predicted to be severely reduced, or totally abolished, in strains with polar plasmid insertions in *psaA* (or further upstream in the operon).

only to relieve inhibition of LytA by TA (Díaz *et al.*, 1996, *Mol Microbiol* **19:** 667–681; Severin *et al.*, 1997, *Microb Drug Res* **3:** 391–400). Ch residues are also essential for surface attachment of a number of ChBPs (Garcia *et al.*, 1999, *Microb Drug Res* **4:** 25–36), including PspA (Yother and White, 1994, *J Bacteriol* **176:** 2976–2985) and most probably LytB. The latter protein is a newly described murein hydrolase that is essential for cell separation (Garcia *et al.*, 1999, *Mol Microbiol*, in press). An attractive hypothesis would then have been that the production of cell wall Ch was affected in the mutants studied by Novak and co-workers. However, this hypothesis was ruled out by immunoblotting with an antibody specific for phosphorylcholine (Novak *et al.*, 1998, *Mol Microbiol* **29:** 1285–1296). In addition, such a hypothesis cannot explain the failure to detect LytA because this protein was not released from the cell in a Ch-independent mutant or when EA was substituted for Ch. This is in contrast to PspA, which requires the presence of Ch residues in the LTA for surface attachment (Yother and White, 1994, *J Bacteriol* **176:** 2976–2985; Yother *et al.*, 1998, *J Bacteriol* **180:** 2093–2101). An alternative hypothesis that deserves further examination is that TA and/or LTA metabolism is affected in these mutants in such a way as to interfere with the attachment of LytA and of the other ChBPs to the pneumococcal cell surface.

Although the Mn requirement for growth of *psa* mutants confirmed our findings (Dintilhac *et al.*, 1997, *Mol Microbiol* **25:** 727–740), we felt concerned by phenotypes (iii) and (iv) because during our investigations we did not notice any autolysis defect in a *psaA* mutant. Therefore, the kinetics of DOC-triggered autolysis of our *psaA* mutant strain were reinvestigated. They appeared indistinguishable from that of the isogenic wild-type parent (Fig. 2, left). Western blot analysis was then performed using antibody to the LytA protein. Similar amounts of LytA were detected in the *psaA* mutant and in the parent strain (data not shown, but see below). Because the genetic background of our mutant differed slightly from that of the *psa* mutants of Novak and co-workers (CP compared with R6 parental strain), we transferred the *psaA* mutation to the latter strain. Normal DOC-triggered autolysis was also observed in the R6 mutant derivative (data not shown).

Susceptibility to autolysis, chain length and penicillin tolerance was also examined in derivatives of the *S. pneumoniae* type 2 strain D39 carrying insertion–duplication mutations (IDM) at three places in the *psa* operon. Strains PsaA$^-$ and ORF3 contained mutations in *psaA* and *psaD* (Fig. 1), respectively, whereas in strain PsaA$^+$ the mutagenesis vector was inserted between the *psaA* and *psaD* open reading frames (Berry and Paton, 1996, *Infect Immun* **64:** 5255–5262). There was no difference in the rate of autolysis in the presence of 0.05% DOC between the wild-type D39 and any of the three mutants. Moreover, there was no apparent difference in the mean chain length in stationary-phase cultures of the four strains ($\approx$6–8 cells in each case; results not shown). Lysates of D39 and the three mutants were also examined by Western blot using polyclonal anti-LytA. The intensity of the 36 kDa immunoreactive band that co-migrates with purified LytA was similar for the D39, PsaA$^-$, PsaA$^+$ and ORF3 lysates, but absent in the lysate of a derivative of D39 with an insertion–duplication mutation in *lytA* (Fig. 2, right). Finally, we tested D39, PsaA$^-$, PsaA$^+$ and ORF3 for penicillin tolerance. At a dose of 0.2 μg ml^{-1} penicillin G (10 times the MIC of D39) over a 6 h period, there was no difference in the rate of penicillin-induced lysis or killing, as judged by decrease in absorbance at 600 nm and viable count, respectively, between any of the four strains (result not shown). There was also no difference in the rate of penicillin-induced lysis in D39 cultures grown in the presence or absence of a 1:100 dilution of polyclonal anti-PsaA.

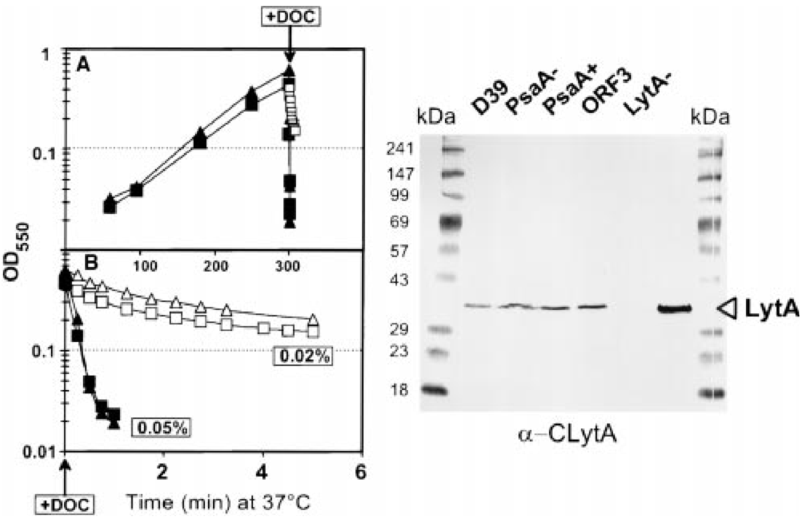


Fig. 2. Kinetics of DOC-triggered autolysis (left) and Western blot analysis of LytA (right) in the *psaA* mutant and in the parent strain (*wt*).
Left. DOC was added (arrow) to *psaA* mutant (■) and *wt* (▲) cultures in the exponential phase of growth in C medium (A). Comparison of the kinetics on an expanded time scale is shown in B. DOC concentration: 0.02% (filled symbols) or 0.05% (open symbols).
Right. Western blot analysis of lysates of D39 and derivatives subjected to SDS–PAGE (12%) and probed using polyclonal mouse anti-LytA. Lanes (from left to right): 1, prestained molecular size markers (band sizes are 241, 147, 99, 69, 57, 43, 29, 23 and 18 kDa from top to bottom); 2, D39; 3, PsaA$^-$; 4, PsaA$^+$; 5, ORF3; 6, LytA$^-$; 7, purified LytA; 8, molecular size markers.

Collectively, all these observations suggest that the use of PsaA in a pneumococcal vaccine formulation, which was questioned on the basis of the possible promotion of penicillin tolerance (Novak *et al*., 1998, *Mol Microbiol* **29:** 1285–1296), should still be considered.

How can we account for the conflicting observations? All *psaA* mutants were generated by IDM, but using different *psaA*-targeting fragments (Fig. 1). Although the truncated *psaA* gene in the CP and D39 mutants is about 260 nucleotides longer than in the mutant of Novak and co-workers, the CP mutation resulted in complete loss of the PsaA protein, as observed by Western blot analysis (Dintilhac *et al*., 1997, *Mol Microbiol* **25:** 727–740). Different non-replicative plasmids were used to construct the *psaA* mutants, pBluescript (Dintilhac *et al*., 1997, *Mol Microbiol* **25:** 727–740), pVA891 (Berry and Paton, 1996, *Infect Immun* **64:** 5255–5262) or pJDC9 (Novak *et al*., 1998, *Mol Microbiol* **29:** 1285–1296). Interestingly, unlike the other two plasmids, pJDC9 contains strong transcriptional terminators (Chen and Morrison, 1987, *Gene* **55:** 179–187). It is, therefore, possible that insertion of pJDC9 to generate the *psaBCA* mutants resulted in stronger polar effects on *psaD* expression than in the other mutants (see legend to Fig. 1). However, 'silencing' of *psaD* is unlikely to be the explanation for the pleiotropic phenotype of the various *psa* mutants of Novak and co-workers because the phenotype of the D39 *psaD* (ORF3) mutant was indistinguishable from D39. At the moment, no explanation(s) that could satisfactorily account for the discrepancy can be proposed. Nevertheless, as stated above, induction of pleiotropic effects including autolysis defects and penicillin tolerance is clearly not a universal feature of *psa* mutants, and therefore PsaA remains a potential pneumococcal vaccine target worthy of careful consideration.

Jean-Pierre Claverys,[1]* Chantal Granadel,[1] Anne M. Berry[2] and James C. Paton[2]
[1] *Laboratoire de Microbiologie et Génétique Moléculaire CNRS-UPR 9007, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France.*
[2] *Molecular Microbiology Unit, Women's and Children's Hospital, North Adelaide, SA 5006, Australia.*
**For correspondence. E-mail claverys@ibcg.biotoul.fr; Tel. (+33) 561 335 911; Fax (+33) 561 335 886.*
Received 25 January, 1999; revised 27 January, 1999; accepted 1 February, 1999.

Identification of putative chromosomal origins of replication in Archaea

Sir,

The mechanisms of DNA replication initiation are quite different in bacteria and eukarya. In bacteria, initiation occurs at a single locus, *oriC*, and is triggered by a single protein, DnaA (Kornberg and Baker, 1992, *DNA Replication*. New York: W. H. Freeman and Co.), whereas, in eukarya, initiation takes place at multiple replication origins that are permanently occupied by origin of replication complexes (ORC) made up of five or six protein subunits. These ORCs are made competent for initiation by the loading of minichromosome maintenance proteins (MCM) (Kearsey and Labib, 1998, *Biochim Biophys Acta* **1398:** 113–136; Pasero and Gasser, 1998, *Curr Opin Cell Biol* **10:** 304–310), an association that is triggered by the protein Cdc6 (Liang and Stillman, 1997, *Genes Dev* **11:** 3375–3386). Until recently, nothing was known about the initiation of DNA replication in the third domain of life, the Archaea, not even whether they have single or multiple replication origins (Edgell and Doolittle, 1997, *Cell* **89:** 995–998). This situation is changing with the advent of archaeal genomics. In particular, Cdc6/Orc1 and MCM homologues have been detected in completely sequenced archaeal genomes (Bernander, 1998, *Mol Microbiol* **4:** 955–961).

In silico attempts have recently been made to identify replication origins in archaeal chromosomes (Grigoriev, 1998, *Nucleic Acids Res* **26:** 2286–2290; Salzberg *et al*., 1998, *Gene* **217:** 57–67). First, in some bacteria, the leading strand contains more G than C, so that the origin (*oriC*) and terminus (*terC*) of chromosome replication can be detected by plotting this GC skew along the genome (Lobry, 1996, *Mol Biol Evol* **13:** 660–665). Grigoriev (1998, *Nucleic Acids Res* **26:** 2286–2290) improved this method by the use of cumulative diagrams that display two peaks when there is a unique origin of replication. Among Archaea, cumulative GC skew diagrams suggested a single *oriC* only for *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*. In *M. thermoautotrophicum*, Grigoriev noticed a homologue of the bacterial chromosome partition gene *soj* close to one of the two peaks, but failed to identify chromosome regions bearing consensus sequences for potential replication origins. Second, some oligomers also appear to have a skewed distribution along the genome. Salzberg *et al*. (1998, *Gene* **217:** 57–67) have located the origin of replication by maximizing the overall oligomer skew on half genomes. In *M. thermoautotrophicum*, Salzberg and co-workers identified the same region of potential replication origin as Grigoriev and noticed the presence of an archaeal homologue of the eukaryotic DNA replication initiator gene *cdc6/orc1* at 5 kb from this putative origin, but they failed to identify skewed oligomers in other Archaea.

We have applied the cumulative skew technique to oligomers from two to eight nucleotides (words) for all completely sequenced prokaryotic genomes. Cumulative word skews give more precise diagrams than GC skews (Fig. 1). An explanation could be that when genome rearrangements

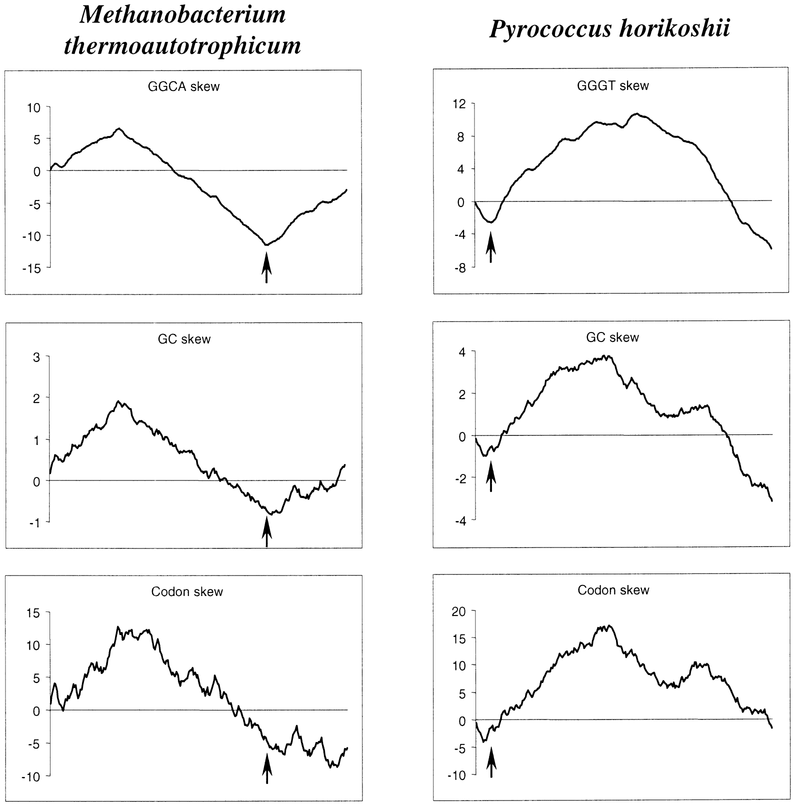


Fig. 1. Cumulative word, GC and codon skew diagrams for *Methanobacterium thermoautotrophicum* and *Pyrococcus horikoshii.* For every word (W), non-overlapping occurrences of W and of its reverse complement W^t (e.g. AATCG for CGATT) were counted on a sliding window along the genome. The location of the window was incremented by 1/240th of the genome for better precision, and its size was 1/50th of the genome for better smoothing, conditions similar to those described previously by Grigoriev (1998, *Nucleic Acids Res* **26**: 2286–2290). Cumulative skew diagrams were obtained by integrating $(nW - nW^t)/(nW + nW^t)$ over the 240 locations, in which *n*W is the number of occurrences of W. Best words were then selected on the smoothness of their cumulative diagram. Our analyses have shown that words of four nucleotides were sufficient for conclusive results. Similarly, W is assumed to be nucleotide G for GC skew (W^t is C). For codon skew, *n*W is the number of codons in the sliding window that are transcribed in the arbitrary positive sense. Ordinates are arbitrary units (cumulated skews), abscissas represent the position in the genome (start is given by complete genome sequences). The arrows indicate the position of Cdc6/Orc1 homologues MT1412 and PH0124 in the complete genome sequences. Genomes were obtained from Smith *et al.* (1997, *J Bacteriol* **179**: 7135–7155) and Kawarabayasi *et al.* (1998, *DNA Res* **5**: 55–76).

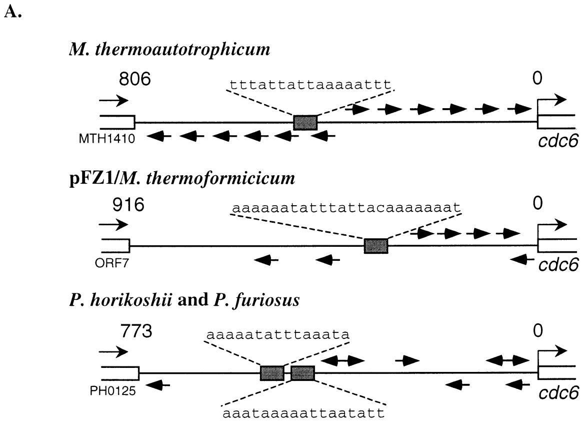
occur the best word skew tends to be restored more rapidly than the GC skew because the latter is probably due to a mutational bias, which is slow because it is selectively neutral. In contrast, the best word skew can be related to the preferential location on the leading strand of signals that are under selective pressure, such as primase recognition sites (Blattner *et al.*, 1997, *Science* **277:** 1453–1462). This technique allowed us to find the origin of replication even when the cumulative GC skew was not informative (*Aquifex aeolicus*, data not shown).

We obtained two-peaked diagrams for the archaea *Pyrococcus horikoshii* (best word GGGT) and *M. thermoautotrophicum* (best word GGCA) (Fig. 1). In the latter case, the peaks were slightly different from those previously found by Grigoriev (1998, *Nucleic Acids Res* **26:** 2286–2290). Because cumulative word skews clearly divide the genomes of *M. thermoautotrophicum* and *P. horikoshii* into two halves, these peaks were promising candidates for origin/termination of bidirectional replication. However, neither GC nor cumulative word skews allow determination of origins for the archaea *M. jannaschii* or *Archaeoglobus fulgidus*, or for the bacterium *Synechocystis*.

Considering that initiator genes are very often located close to the origin in bacteria, plasmids and viruses, we looked at the ORFs located in the regions surrounding the peaks. Most interestingly, these regions contained archaeal homologues of *cdc6/orc1* in *M. thermoautotrophicum* (accession number MT1412) and in *P. horikoshii* (accession number PH0124) (Fig. 1). In contrast, the *soj* gene previously noticed by Grigoriev (1998, *Nucleic Acids Res* **26:** 2286–2290) is located 40 kb and 550 kb away from the *cdc6/orc1*-containing peaks of *M. thermoautotrophicum* and *P. horikoshii* respectively.

As observed at the origin of replication in most bacteria, the plot of $G - C/G + C$ shifted from negative to positive at the peak close to the *cdc6/orc1* gene, in agreement with this location being the origin. Moreover, all rRNA genes and the majority of ribosomal protein genes were transcribed in the same direction as DNA replication under our working hypothesis (not shown). Such organization is expected to avoid head-to-head collisions between transcription complexes and replication forks (French, 1992, *Science* **258:** 1362–1365). Indeed, using cumulative diagrams again, the codon skew correlated reasonably well with the cumulative word skew both in *M. thermoautotrophicum* and *P. horikoshii* (Fig. 1).

The congruence between the location of a putative initiator gene and GC, oligomer and codon skews prompted us to look for putative origin sequences surrounding the



B.

Organisms	Repeated nucleotides	Times present
M.t.	tta cac tt- gaa at	8
	tta tac tt- gaa gg	1
	tta cag tt- gaa ag	1
	tta cac tt- gaa ac	1
	tta cag tt- gaa at	1
pFZ1	tta caa gta gaa at	2
	tta caa ttc gca at	2
	ttt caa tta gaa at	1
	tta cat ata gaa at	1
	tta caa ttc gaa ac	1
P. h. and *P. f.*	ttc cag tg- gaa at	3
	ctc cag tg- gaa at	3
	ttc caa ag- gaa at	1
	ttc cac tg- gaa ct	1
	ttc cac tg- gaa at	1
	ttc cag tg- gaa gt	1
	ctc cac tg- gaa ac	1
	ctc cac ag- gaa at	1
	ctc cat tg- gaa at	1
Consensus	tTa CAg Tg- GAA AT c c c	

Fig. 2. Schematic physical maps of *oriC* and alignments of the repeats.
A. Repeats similar in the three *oriC* are represented by black arrows whose direction indicates sense. Longer direct repeats are present in *M. thermoautotrophicum* and pFZ1 *oriC,* but have not been represented for simplicity. Grey boxes represent central AT-rich elements whose sequences are listed.
B. Repeats are aligned for the three *oriC* we identified, and displayed with their number of occurrences. A consensus sequence is shown. *M. t.*, *Methanobacterium thermoautotrophicum*; pFZ1, plasmid pFZ1 from *Methanobacterium thermoformicicum*; *P. h.*, *Pyrococcus horikoshii*; *P. f.*, *Pyrococcus furiosus*. Accession number for pFZ1 is GenBank X67212.

cdc6/orc1 gene. Replication origins are usually located in large intergenic regions that exhibit complex patterns of AT-rich elements as well as direct and inverted repeats (Kornberg and Baker, 1992, *DNA Replication*, New York: W. H. Freeman and Co.). Indeed, we found such regions at the 5′ end of the two archaeal *cdc6/orc1* genes identified by the word skew diagrams (Fig. 2). In *M. thermoautotrophicum*, this region contains 12 copies of a 13 bp repeat that are symmetrically distributed around an internal AT-rich element (Fig. 2). The spacing of these repeats is remarkably regular (about 50 bp), and five of them are included in larger perfect repeats of 25 and 21 bp respectively. The corresponding region of *P. horikoshii* also contains repeats of 13 bp distributed on either side of two central AT-rich elements. These *Pyrococcus* repeats turned out to be strikingly similar to those detected in *M. thermoautotrophicum* (Fig. 2). As expected for essential regulatory elements, the repeats and AT-rich elements of the putative *P. horikoshii oriC* were conserved in *Pyrococcus furiosus* (Fig. 2 and data not shown).

Interestingly, an archaeal homologue of *cdc6/orc1* has been previously detected in the plasmid pFZ1 from *Methanobacterium thermoformicicum* (ORF1, SWISS-PROT P29570) (Edgell and Doolittle, 1997, *Cell* **89:** 995–998). We checked whether repeats similar to those identified in the putative archaeal *oriC* were also present in pFZ1. Indeed we found seven copies of a 14 bp repeat that matches very well with archaeal *oriC* consensus sequences (Fig. 2). They are again present in an intergenic AT-rich region at the 5′ end of the *cdc6/orc1* gene. This suggested that an archaeal chromosomal replication origin is used by pFZ1 for its own replication.

The presence of both these sequences and a *cdc6/orc1* homologue on a plasmid suggests that Cdc6/Orc1 itself recognizes the repeated sequences found in *M. thermoautotrophicum*, *Pyrococcus* and pFZ1 origins. The latter notion can be further supported by recent observations indicating that ORC1, ORC4 and ORC5 subunits of the ORC complex all appear to be related to each other (Tugal *et al*., 1998, *J Biol Chem* **273:** 32421–32429), and that all of these subunits are known to interact with ARS sequences in yeast (Lee and Bell, 1997, *Mol Cell Biol* **17:** 7159–7168). Therefore, the function(s) of archaeal Cdc6/Orc1 homologues could be more similar to that of Orc1, although they are not more similar in sequence to Orc1 than to Cdc6 eukaryotic proteins.

In *P. furiosus*, the *cdc6/orc1* gene is co-transcribed with the genes encoding the subunits DP1 and DP2 of a recently identified archaeal specific DNA polymerase (Uemori *et al*., 1997, *Genes Cells* **2:** 499–512). The promoter region thus defined overlaps with the *oriC* sequences identified in the two *Pyrococcus* (not shown), suggesting an interplay between control of DNA replication initiation and transcription of initiator and replicator genes. Interestingly, the *M. thermoautotrophicum DP1* gene (accession number MT1405) is located only 5 kb away from the putative *oriC*. This could facilitate the formation of the replication complex at the origin by promoting a direct interaction between archaeal Cdc6/Orc1 and DP1.

Altogether, our results strongly suggest that we have identified the unique *oriC* of *M. thermoautotrophicum, P. horikoshii* and *P. furiosus*, as well as consensus sequences recognized by initiator protein(s). Considering the conservation of these features between *Pyrococcus* species and *M. thermoautotrophicum,* it might seem surprising that we failed to identify putative *oriC* in *A. fulgidus* and *M. jannaschii*. One possible explanation is that these archaea use multiple origins instead of a single *oriC*. However, we suppose that all prokaryotes have a single origin of replication but that the *in silico* approach can be confused by very frequent genome rearrangements and/or by intrinsic properties of the genome, such as mutational bias or primase recognition site (to be detailed elsewhere). Surprisingly, the genome of *M. jannaschii* does not contain a Cdc6/Orc1 homologue, but four homologues of MCM instead of only one in other Archaea (Bernander, 1998, *Mol Microbiol* **4:** 955–961). This suggests that the apparatus for replication initiation is more flexible than often expected and can be affected by non-orthologous displacement. In agreement with this conclusion, it has been recently shown than disruption of the *dnaA* gene *Synechocystis* has no phenotypic effect (Richter *et al*., 1998, *J Bacteriol* **180:** 4946–4949), indicating that another protein must be the replication initiator in this bacterium.

This opens new perspectives for studying archaea. For example, archaeal minichromosomes resembling pFZ1 could be the starting point for the construction of cloning vectors. They might be used to analyse the mechanisms of DNA replication initiation, elongation and cell cycle regulation *in vitro*. Such studies may have important consequences for our understanding of these processes in eukaryotes.

Acknowledgements

This work was supported by research grants from the European Union Biotechnology programme (BIO4-CT 96-0488) and by the Association de Recherche contre le Cancer. H.M. is supported by a Marie Curie postdoctoral training grant from the European Union.

Philippe Lopez,[1] Hervé Philippe,[1]* Hannu Myllykallio[2] and Patrick Forterre[2]
[1] *Phylogénie et Evolution Moléculaires, UPRESA Q8080, Bat 444, Université Paris-Sud, 91405 Orsay Cedex, France.*
[2] *Institut de Génétique et Microbiologie, Bat 409, Université Paris-Sud, 91405 Orsay Cedex, France.*
**For correspondence. E-mail Herve.Philippe@bc4.u-psud.fr; Tel. (+33) 1 69 15 64 81; Fax (+33) 1 69 15 68 03*
Received 14 January, 1999; revised 28 January, 1999; accepted 1 February, 1999.

Whole-genome sequence annotation: 'Going wrong with confidence'

Sir,

The seventeenth complete genome, that of *Chlamydia trachomatis*, has been published recently, infusing the public database with another 894 protein sequences (Stephens *et al*., 1998, *Science* **282**: 754–759). The annotation of this genome is different from the previous 16 analyses, in that it introduces a number of spurious functional assignments based on questionable predictions (Table 1).

First, new terms have been used that are semantically meaningless, for example 'predicted' or 'possible' functions (e.g. CT149, see Table 1 for identifiers), without explicit reference to the prediction methods used. Secondly, the presence of motifs and/or domains has been widely used as a substitute for function assignments (e.g. CT555). Thirdly, there are a number of overpredictions, i.e. overly specific assignments with no, or very remote, similarity to proteins of known function (e.g. CT775).

All the above cases have the potential for resulting in a significant error propagation effect, especially when they refer to large hypothetical families from previous and ongoing genome projects. In other words, the *Chlamydia* genome is an actual snapshot of error propagation in public sequence databases (see also Karp, 1998, *Bioinformatics* **14:** 753–754). Thus, some caution should be exercised when assigning new functions based exclusively on the annotations of this genome.

Overambitious annotation projects have appeared ever since the first genome sequences started covering unknown territory with novel protein families (Casari *et al*., 1995, *Nature* **376**: 647–648). For elusive cases, conflicting assignments are common and are inherent in the process of function prediction. Yet, there appears to exist a limit, above which predictions do more harm than good.

A unique example of annotation abuse is the study of the *Methanococcus jannaschii* genome by Koonin and colleagues (http://www.ncbi.nlm.nih.gov/Complete_Genomes/MJan2/mjtable.html) (Koonin *et al*., 1997, *Mol Microbiol* **25**: 619–637). While three independent analyses have not surpassed a 50% level of function prediction (with 95% agreement; Andrade *et al*., 1997, *Comput Appl Biosci* **13**: 481–483), the study by Koonin *et al*. claimed that 'about 70% of the archaeal proteins were predicted with varying precision' (Koonin *et al*., 1997, *Mol Microbiol* **25**: 619–637). In this additional 20%, there are cases such as MJ0539, which was predicted to be a cysteinyl-tRNA synthetase, later found to be the lysyl-tRNA synthetase (Ibba *et al*., 1997, *Science* **278**: 1119–1122) (more examples in Table 1). Further to the above-mentioned three sources of error, there also exist cases apparently inconsistent with the clusters of orthologous genes (COGs), described previously by the same group (Tatusov *et al*., 1997, *Science* **278**: 631–637).

Putative assignments based on 'twilight-zone' similarities and/or the mere presence of motifs should only serve as working hypotheses in search of a function. As such, they can reflect the current status and wealth of

Table 1. A comparison of 30 selected cases of potential genome sequence annotation conflicts between published predictions and our analysis.

ORF identifier	Erroneous prediction	Reason	Current status
Chlamydia			
CT149	Possible hydrolase	O	Hypothetical (similar to CT206)
CT206	(Predicted acyltransferase family)	O	Hypothetical (similar to CT149)
CT312	Predicted ferredoxin	F	Unique
CT422	Possible metalloenzyme	F	Hypothetical
CT498	FAD-dependent oxidoreductase	U	*gidA* family (predicted oxidoreductases)
CT555	SWI/SNF family helicase	O	SWI/SNF family (C-terminal domain only)
CT738	Metal-dependent hydrolase	F/O	Hypothetical
CT775	snGlycerol 3-P acyltransferase	O	Remote similarity to snG-3-P acyltransferases
CT804	Predicted kinase	U	Remote similarity to homoserine kinase family
CT859	Metalloprotease	F	LytB family
Methanococcus			
MJ0018	Flagellin	O	Unique
MJ0028	Hydrogenase expression factor	F	Thiamine monophosphate kinase
MJ0037	Phosphoesterase	D	Hypothetical
MJ0039	Ribosomal protein L7/L12 paralogue	F	Hypothetical
MJ0047	Zn-dependent hydrolase	F	Cleavage and polyadenylation specificity factor
MJ0068	Thioredoxin-like protein	F	Hypothetical
MJ0094	Methyl coenzyme M reductase operon protein homologue	F/O	Hypothetical
MJ0096	Putative permease	F	Unique
MJ0107	Dihydropteroate pyrophosphorylase (dhps)	O	dhps-containing protein (central domain)
MJ0123	Putative Zn-dependent protease	F	Hypothetical
MJ0287	HTH transcription regulator	F	Hypothetical
MJ0290	HTH transcription regulator	F	Hypothetical
MJ0296	Zn-dependent hydrolase	F	Hypothetical
MJ0306	Ferredoxin	F	Unique
MJ0416	Rubrerythrin	F	Unique
MJ0418	Periplasmic protein	F	Hypothetical
MJ0431	Integral membrane protein	F	Unique
MJ0464	Translation initiation factor IF1	O	Hypothetical
MJ0539	Cysteinyl-tRNA synthetase	F	Lysyl-tRNA synthetase
MJ0951	Nucleotidyltransferase, glycerol-phosphate cytidyltransferase	D	Hypothetical

Columns: ORF identifier: *C. trachomatis* (CT) and *M. jannaschii* (MJ) ORF sequences; Erroneous prediction: published predictions for the corresponding genes; Reason: F, false/no evidence; O, overprediction; U, underprediction; D, disagreement with COGs. Current status (http://geta.life.uiuc.edu/~nikos/Mjannotations.html mirrored at http://www.ebi.ac.uk/research/cgg/annotation/MJannotations.html): our own results from manual analysis (identical search methods used against the NRDB at the EBI: 343 038 sequences, 105 533 698 residues, 11 December 1998). Unique signifies no homologues in the database, while hypothetical means the presence of homologous proteins without specific function assignments.

our collective knowledge and should therefore be continuously analysed within species-specific genome databases before their final admission to the public database. A concerted effort using continuous curation by experts and clear demonstration of evidence is therefore necessary to avoid this serious error propagation effect in computational genomics.

Nikos C. Kyrpides[1] and Christos A. Ouzounis[2]*
[1]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
[2]Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.
**For correspondence. E-mail ouzounis@ebi.ac.uk; Tel. (+44) 01223 494653; Fax (+44) 01223 494468.*
Received 23 January, 1999; revised 8 February, 1999; accepted 10 February, 1999.

The *Escherichia coli* ABC transporters: an update

Sir,

A recent MicroGenomics review article presented the ATP-binding cassette (ABC) proteins encoded within the *Escherichia coli* genome (Linton and Higgins, 1998, *Mol Microbiol* **28:** 5–13). We have also analysed the ABC transporters found universally in living organisms (see Saier, 1998, In *Advances in Microbiological Physiology*. Poole, R.K. (ed.) **40:** 81–136). Our most recent analyses led us to update the descriptions in the article of Linton and Higgins. These descriptions are completed or corrected below, and we provide further interesting information regarding some of the relevant proteins (see Table 1 in the article by Linton and Higgins).

The following systems are probably exporters instead of importers: (i) f583 is most closely related to the DrrA protein of *Streptomyces peucetius*, a constituent of a known

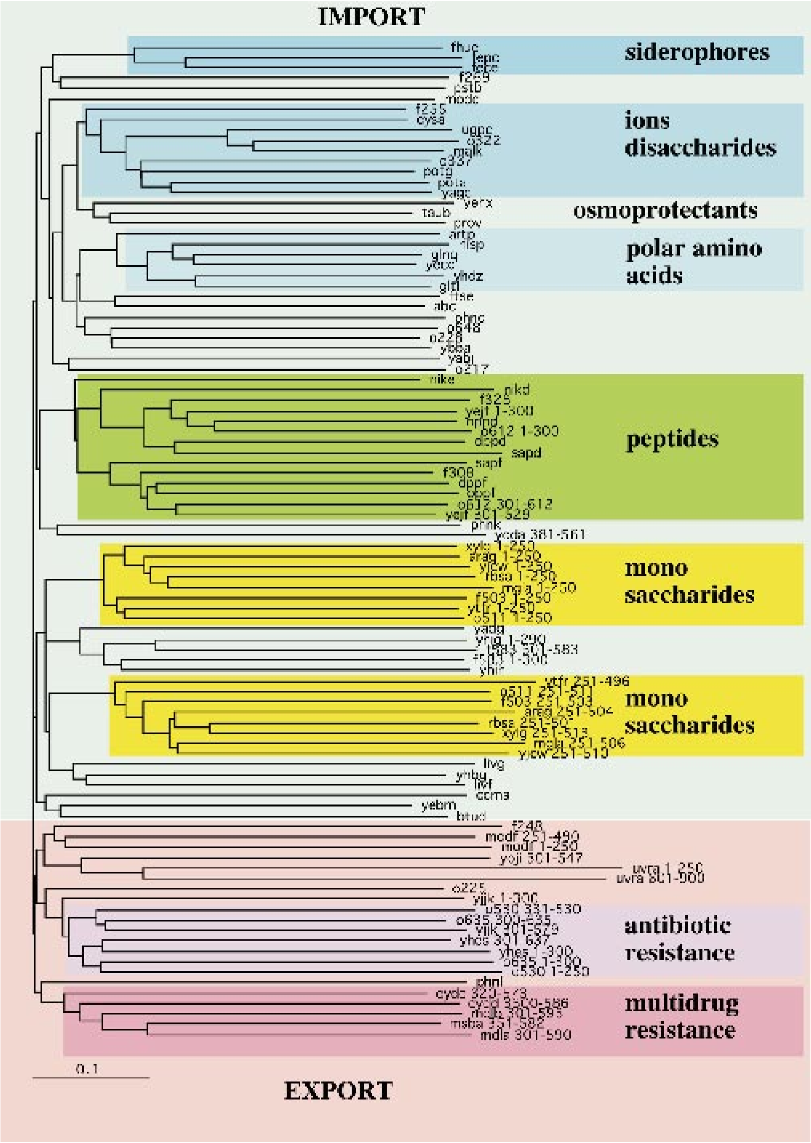


Fig. 1. Phylogenetic tree of *Escherichia coli* ATP-binding proteins. Groups of proteins from systems with known or predicted specificity are represented within coloured rectangles. The upper part of the figure contains all known import systems and the lower part putative export systems.

drug efflux pump. Homologues of DrrA are associated with one or two inner membrane proteins (f377 and f368 in the case of f583) and a protein homologous to membrane fusion proteins (MFPs, Dinh *et al*., 1994, *J Bacteriol* **176:** 3825–3831) present in many export systems (f332 in the case of f583). (ii) YhiG and YhiH exhibit striking sequence identity to f583, f377 and f368. YhiI would be the associated MFP. (iii) o648 is homologous to o228 and YbbA, and homologues are found in many bacterial genomes. They are accompanied by one or two hydrophobic membrane proteins, which, in the case of o648, is fused to the ATP-binding protein. These are sometimes found encoded in operons with genes encoding proteins having strong similarity to members of the MFP family.

YcbE could be part of an operon encoding f278 (a conserved inner membrane protein) and f333 (a putative substrate-binding protein). This system is therefore probably an importer and not an exporter. The functions of the following two systems are known: (i) YebM is similar to proteins involved in the import of Mn^{2+}, Zn^{2+} and Fe^{2+}. YebM is associated with YebI (a conserved inner membrane protein) and YebL (a putative substrate-binding protein). This system has been shown to catalyse high-affinity zinc uptake (Patzer and Hantke, 1998, *Mol Microbiol* **28:** 1199–1210). (ii) The YjcVWX operon is known to function in the utilization of D-allose (Kim *et al*., 1997, *J Bacteriol* **179:** 7631–7637).

Some known or putative constituents of ABC systems were not described in the Linton and Higgins review as follows: (i) the CysATW (sulphate–thiosulphate) system functions with two distinct periplasmic binding receptors, CysP (thiosulphate uptake) and SbpA (sulphate uptake).

(ii) The LivFGHM (hydrophobic amino acid) system functions with two distinct periplasmic binding receptors, LivJ (leucine/isoleucine/valine receptor) and LivK (leucine-specific receptor). (iii) Homologues of f248 (cited as f284 in Table 1 of Linton and Higgins) are found in several bacterial and chloroplast genomes, and they are always associated with a conserved protein. In *E. coli* two such proteins (f508 and f423) flank the *f248* gene. (iv) o322 is an ATP-binding protein most similar in sequence to those in the oligosaccharide ion family (family 5). It is probably part of an operon where o430 encodes the substrate-binding protein and o293 and o280 encode the cytoplasmic membrane proteins. This system is therefore most likely to be an oligosaccharide uptake system. (v) YecC belongs to the polar amino acid transporter family (family 4). It is associated with YecS (a conserved cytoplasmic membrane protein) and FliY (a putative substrate-binding protein). FliY has been identified as the cystine-binding protein, and consequently this permease probably transports this amino acid as well as diaminopimelate (B. Schneider, C. Furlong and M. H. Saier, Jr, unpublished observations). (vi) f535 is an orphan periplasmic-binding receptor that clusters in the peptide-binding receptor family.

It is possible to predict the functions of the following systems. (i) YagC could be part of an operon that has been disrupted by an insertion sequence (IS) element. Partial open reading frames (ORFs) around this IS element encode protein fragments that display strong similarity to cytoplasmic membrane proteins of the oligosaccharide ion family (family 5). (ii) o530 is an ATP-binding protein homologous to the yeast GCN-20 protein, a protein involved in the regulation of translational initiation. Based on the sequence similarity observed, a comparable function in *E. coli* can be proposed.

With respect to the phylogenetic characterization of ABC transporter protein constituents, we have provided evidence for the co-evolution of the constituents of ABC-type uptake systems from a common evolutionary origin with minimal shuffling (Tam and Saier, 1993, *Microbiol Rev* **57:** 320–346; Saurin and Dassa, 1994, *Prot Sci* **3:** 325–344; Kuan *et al*., 1995, *Res Microbiol* **146:** 271–278). One of us (E. Dassa) devised a tree (Fig. 1) of *E. coli* ATP-binding proteins using the same method as that used by Linton and Higgins (1998, ibid.). This tree proved to be similar to that reported in this paper, but since it is unrooted, we have chosen a representation that highlights the fact that exporters and importers segregated early, as documented (Saurin *et al*., 1999, *J Mol Evol* **48:** 22–41). This representation suggests that import and export systems separated very early during the evolution of the ABC superfamily, before divergence of the proteins that comprise either of these two groups.

A complete list of the ABC and other transporters in *E. coli* and in various other organisms (Paulsen *et al*., 1998a, *J Mol Biol* **277:** 573–592; Paulsen *et al*., 1998b, *FEBS Lett* **430:** 116–125) is available on the World Wide Web (http://www-biology.ucsd.edu/~ipaulsen/transport/titlepage.html), and the classification of the ABC superfamily into 44 families as well as the classification of 200 other transporter families is described in a separate web site (http://www-biology.ucsd.edu/~msaier/transport/titlepage.html).

Elie Dassa,[1]* Maurice Hofnung,[1] Ian T. Paulsen[2] and Milton H. Saier Jr[2]
[1]*Unité de Programmation Moléculaire et Toxicologie génétique, CNRS URA 1444, Institut Pasteur, 25, Rue du Dr Roux, F75724 Paris Cedex 15, France.*
[2]*Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA.*
**For correspondence. E-mail elidassa@pasteur.fr; Tel. 33 1 45 68 88 31; Fax 33 1 45 68 88 34.*

Received 10 February, 1999; revised 16 February, 1999; accepted 18 February, 1999.

The *ipdC* promoter auxin-responsive element of *Azospirillum brasilense*, a prokaryotic ancestral form of the plant AuxRE?

Sir,

Auxins are a major class of plant growth regulators known to be involved in diverse processes at the whole plant level and at the cellular level. Regulation of these processes by auxin is thought to be the result of modified gene expression.

A number of early auxin-induced plant genes have recently been cloned and characterized (Napier and Venis, 1995, *New Phytol* **129:** 167–201). A common short sequence element TGTCTC (or the degenerate version (G/T)GTCCCAT), termed auxin-responsive element (AuxRE), has been identified in the promoters of some of these auxin-regulated genes, including the soybean *GH3* and *SAUR 15A* genes, and the pea *PS-IAA4/5* gene (Abel *et al*., 1996, *BioEssays* **18:** 647–654; Guilfoyle *et al*., 1998, *Plant Physiol* **118:** 341–347). In naturally occurring auxin-responsive promoters, these AuxREs have been found to function with a coupling element overlapping with or adjacent to the TGTCTC half-site. Within these composite *cis*-acting elements, the coupling element confers tissue-specific or development-specific expression to the auxin-regulated promoter, and the TGTCTC half-site acts to repress this expression when auxin levels are low. Derepression then follows increasing auxin levels. Experiments with synthetic AuxREs have shown that direct repeats or palindromes of the conserved half-site are preferred in conferring high inducibility by auxin to a

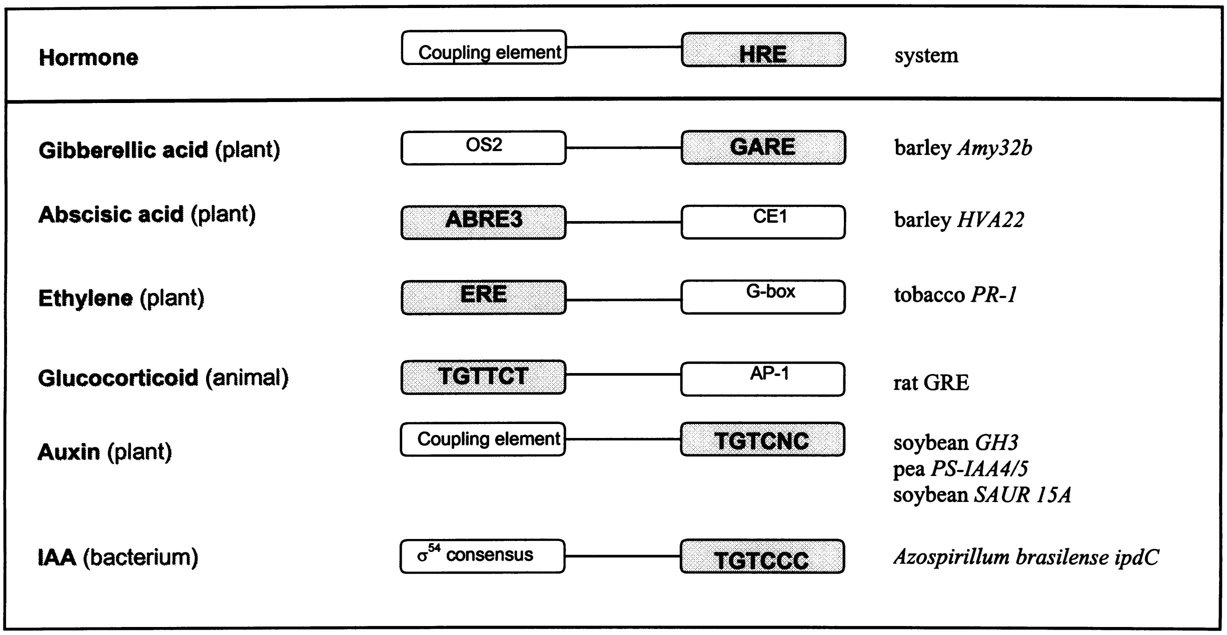


Fig. 1. Comparison of plant and animal hormone-responsive elements to the auxin-responsive *ipdC* sequence element of *Azospirillum brasilense*. HRE, hormone response element; OS2, *opaque-2* DNA binding site; GARE, gibberellic acid response element; ABRE, abscisic acid response element; CE1, coupling element; ERE, ethylene response element; AP-1, activator protein-1.

minimal promoter construct (Ulmasov *et al.*, 1997, *Plant Cell* **9:** 1963–1971). A TGTCTC palindrome was subsequently identified in the pea *PS-IAA4/5* promoter. The 'coupling model' of plant hormone response complexes, in which general transcriptional regulators are adjacent to sites that confer hormone inducibility, is valid for most plant hormones, e.g. gibberellin (Lanahan *et al.*, 1992, *Plant Cell* **4:** 203–211), abscisic acid (Shen and Ho, 1995, *Plant Cell* **7:** 295–307) and ethylene (Mason *et al.*, 1993, *Plant Cell* **5:** 241–251) (Fig. 1). The modular composition of hormone-inducible promoter elements is apparently conserved over the plant and animal kingdoms, as similarities between TGTCTC AuxREs and TGTTCT glucocorticoid or steroid hormone response elements (GREs or HREs) have been noted (Ulmasov *et al.*, 1998, *Science* **276:** 1865–1868). Both contain the conserved response element and a second DNA binding site for an activator protein (Fig. 1).

The biosynthesis of plant growth-regulating substances is not restricted to plants, and biosynthetic pathways have also been identified in many micro-organisms, such as fungi, phytopathogenic and plant growth-stimulating bacteria. *Azospirillum brasilense*, a plant growth-promoting rhizobacterium, produces gibberellins, cytokinins and the auxin indole-3-acetic acid (IAA). IAA biosynthesis in this bacterium occurs through different biosynthetic pathways (Prinsen *et al.*, 1993, *Mol. Plant-Microbe Interact* **6:** 609–615), one of which is the indole-3-pyruvate (IPyA) pathway. In the IPyA pathway, the IPyA decarboxylase is responsible for the conversion of IPyA to indole-3-acetaldehyde, and the *ipdC* gene encoding this enzyme has been cloned and characterized in *A. brasilense* (Costacurta *et al.*, 1994, *Mol Gen Genet* **243:** 463–472). Surprisingly, transcription of the *A. brasilense ipdC* gene, analysed by means of an *ipdC* promoter–*gusA* translational fusion and by Northern analysis, is specifically induced by IAA and by other compounds that are known as synthetic auxins in plant physiology (Vande Broek *et al.*, 1999, *J Bacteriol* **181:** 1338–1342).

Upstream of the *ipdC* gene, a DNA motif can be recognized: a consensus sequence for a σ^{54}-dependent promoter (Fig. 1) that partially overlaps at the 3′ end with a TGTCCC element, reminiscent of the AuxREs in plants. This sequence could thus represent a new type of operator. On the basis of this observation, we propose that the combination of the coupling element (in this case the σ^{54} consensus sequence) and the TGTCCC element, similar to the modular build-up of hormone responsive promoters in plants and animals, is responsible for conferring auxin inducibility to the *ipdC* promoter. In plants, auxin response factors (ARFs) were discovered that bind to AuxREs with the consensus sequence TGTCNC. ARF1 was the first of those transcription factors to be identified in *Arabidopsis* (Ulmasov *et al.*, 1998, ibid.). ARFs are thought to act as repressors or activators depending on the binding of other ARFs, the binding of transcriptional factors for the coupling elements or through interactions with another class of transcriptional regulators, the Aux/IAA proteins. All of these interactions could somehow be mediated by auxin. Auxin-binding proteins and transcription factors remain to be identified in *A. brasilense*. It has already been suggested that a repressor/activator may work by decreasing/increasing the concentration of RNA polymerase at a promoter capable of forming an open complex (Müller-Hill, 1998, *Mol Microbiol* **29:** 13–18). This strategy is called increase in local concentration or recruitment. It could well be that a new type of *A. brasilense* transcription factors, binding at the TGTCCC consensus site, is able to influence the formation of the RNA polymerase complex through synergistic recruitment of additional proteins for transcriptional activation. These interactions could be directly or indirectly mediated by auxin.

If this hypothesis is confirmed experimentally, the molecular mechanism of conferring hormone inducibility to genes in eukaryotes might have originated in prokaryotes.

Acknowledgements

The authors gratefully acknowledge Dr René De Mot for critical reading of the manuscript.

Mark Lambrecht, Ann Vande Broek, Filip Dosselaere and Jos Vanderleyden*
F. A. Janssens Laboratory of Genetics, K.U. Leuven,
Kard. Mercierlaan 92, B-3001 Heverlee, Belgium.

Received 1 February, 1999; revised 16 February, 1999; accepted 19 February, 1999.