Volume 13 Number 15 1985 Nucleic Acids Research

Nucleotide sequence of the *Bacillus subtilis* xylose isomerase gene: extensive homology between the *Bacillus* and *Escherichia coli* enzyme

Martin Wilhelm and Cornelis P.Hollenberg

Institut für Mikrobiologie, Universität Düsseldorf, Universitätsstrasse 1, 4000 Düsseldorf, FRG

Received 2 March 1985; Revised and Accepted 16 July 1985

ABSTRACT

The xylose isomerase gene from Bacillus subtilis was cloned from a genomic BamH1 library by complementation of an isomerase defective Escherichia coli strain as previously described.

The ATG initiation codon is preceded by a Shine-Dalgarnosequence and two hexamers being characteristic for the promoterregion of Bacillus genes. The structural gene consists of 1320 base pairs, thus coding for a polypeptide chain of 440 amino acids with a molecular weight of 49 680. The polypeptide primary structure shows over 50% homology to that of the E. coli xylose isomerase.

INTRODUCTION

We have cloned and analysed the xylose isomerase gene from Bacillus subtilis in order to study its expression in Saccharomyces cerevisiae. In the process of cloning in Escherichia coli we found that the isomerase activity was expressed in E. coli only after IS5 insertion (1). The effect of IS5 insertion could be substituted by the insertion of the E. coli lacUV5 promoter close to the initiation codon of the isomerase structural gene. The xylulokinase gene of B. subtilis is located distal from the isomerase gene (1). The E. coli xylose isomerase gene has been sequenced by two groups (2,3) reporting conflicting data regarding the length of the open reading frame.

In this communication we present the nucleotide sequence of the xylose isomerase gene and its 5' and 3' flanking regions. Furthermore, we have compared the derived amino acid sequence to that reported for the E. coli enzyme and found 50 % homology extending over the entire polypeptide up until the end of the longest published reading frame (2).

MATERIALS AND METHODS

DNA isolation and analysis. Plasmid DNA was isolated by alkaline lysis (4) and purified by CsCl density gradient centrifugation. DNA was cleaved with appropriate enzymes and fractionated on 0.7 % agarose gels.

DNA fragments used for sequencing were 3' ^{32}P labeled with Klenow polymerase according to the method of Maniatis (5) and separated on 1% low-melting agarose gels (6). The fragments were isolated from the gel by melting at 60°C in the presence of an equal volume of phenol and centrifuged in an Eppendorf centrifuge for 10 minutes. The supernatant was extracted once with phenol/chloroform, once with chloroform and the DNA was ethanol precipitated.

DNA sequence analysis. DNA sequencing was performed according to Sanger et al. (7), using the M13 system (8) and according to Maxam and Gilbert (9).

RESULTS

The genes for xylose isomerase and xylulokinase have been cloned on a 5.8 kb BamHI fragment from B. subtilis (1). By subsequent deletion analyses we were able to localize the xylose isomerase gene on a 1.8 kb EcoRI-BglII fragment (1). The strategy for sequencing this fragment is given in Fig.1. The complete nucleotide sequence and the amino acid sequence deduced from the xylose isomerase coding region are shown in Fig.2.

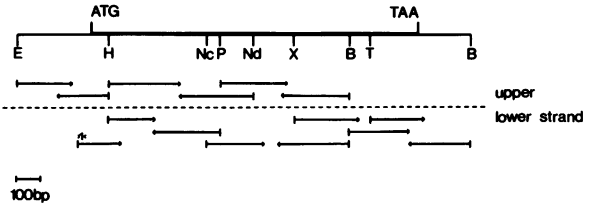

Figure 1 Sequencing strategy for the EcoRI-BglII fragment. The >1< designated arrow marks a sequence read from a Bal31 deletion fragment. The start and the stop codons of the xylose isomerase gene are indicated. Abbreviation for restriction enzymes are: B=BglII; E=EcoRI; H=HindIII; Nc=NcoI; Nd=NdeI; P=PstI; T=TaqI; X=XbaI.

```
  1 GAATTCTTTA CTTTTTTTGA CAGGTTTGAT CATTGCGATA TCCATTATTC ATCATTTGTA TAACCTCTAA ATTAAGTTAA AATTTTTTGT GTTCAGTATG ATTTAGTACA TAGCGAATCT
                                                -132                  -109
121 TACCTTTATT ATATCTAATG TGTTCATGAA AAACTAAAAA AAATATTGAA AATACTGATG AGGTTATTTA AGATTAAAAT AAGTTAGTTT GTTTGGGCAA CAAACTAATG TGCAACTTAC
                                                -14
241 TTACAATATG ACATAAAATG CATCTGTATT TGAATTTATT TTTAAGGAGG AAATAAC
   +1
298 ATG GCT CAA TCT CAT TCT AGT TCA GTT AAC TAT TTT GGA AGC GTA AAC AAA GTG GTT TTC GAA GGG AAA GCT TCC ACT AAT CCT TTA GCA TTT AAA
    Met Ala Gln Ser His Ser Ser Ser Val Asn Tyr Phe Gly Ser Val Asn Lys Val Val Phe Glu Gly Lys Ala Ser Thr Asn Pro Leu Ala Phe Lys

394 TAT TAT AAT CCT CAA GAA GTA ATC GGC GGA AAA ACG ATG AAA GAG CAT TTG CGA TTT TCT ATT GCC TAT TGG CAT ACA TTT ACT GCT GAT GGC ACA
    Tyr Tyr Asn Pro Gln Glu Val Ile Gly Gly Lys Thr Met Lys Glu His Leu Arg Phe Ser Ile Ala Tyr Trp His Thr Phe Thr Ala Asp Gly Thr

490 GAC GTT TTT GGA GCA GCT ACA ATG CAA AGA CCA TGG GAT CAC TAT AAA GGC ATG GAT CTA GCT AGG GCA AGA GTA GAA GCA GCA TTT GAG ATG TTT
    Asp Val Phe Gly Ala Ala Thr Met Gln Arg Pro Trp Asp His Tyr Lys Gly Met Asp Leu Ala Arg Ala Arg Val Glu Ala Ala Phe Glu Met Phe

586 GAA AAA CTA GAT GCA CCA TTT TTT GCT TTT CAT GAT CGA GAT ATT GCA CCA GAA GGA AGT ACG TTA AAA GAG ACA AAT CAA AAT TTA GAT ATT ATC
    Glu Lys Leu Asp Ala Pro Phe Phe Ala Phe His Asp Arg Asp Ile Ala Pro Glu Gly Ser Thr Leu Lys Glu Thr Asn Gln Asn Leu Asp Ile Ile

682 GTG GGC ATG ATT AAG GAT TAC ATG AGA GAT AGC AAC GTT AAG TTA TTA TGG AAT ACT GCA AAC ATG TTT ACG AAC CCC CGT TTC GTC CAT GGA GCC
    Val Gly Met Ile Lys Asp Tyr Met Arg Asp Ser Asn Val Lys Leu Leu Trp Asn Thr Ala Asn Met Phe Thr Asn Pro Arg Phe Val His Gly Ala

778 GCG ACT TCT TGT AAT GCA GAT GTG TTT GCG TAT GCT GCA GCA CAA GTA AAA AAA GGG TTA GAA ACA GCA AAA GAG CTT GGC GCC GAG AAC TAT GTA
    Ala Thr Ser Cys Asn Ala Asp Val Phe Ala Tyr Ala Ala Ala Gln Val Lys Lys Gly Leu Glu Thr Ala Lys Glu Leu Gly Ala Glu Asn Tyr Val

874 TTT TGG GGC GGC CGT GAA GGA TAC GAA ACA TTG TTA AAT ACC GAT TTA AAA TTT GAG CTT GAT AAT TTG GCG AGA TTT ATG CAT ATG GCA GTA GAT
    Phe Trp Gly Gly Arg Glu Gly Tyr Glu Thr Leu Leu Asn Thr Asp Leu Lys Phe Glu Leu Asp Asn Leu Ala Arg Phe Met His Met Ala Val Asp

970 TAT GCG AAG GAA ATC GAG TAT ACA GGG CAG TTT TTG ATT GAA CCA AAA CCA AAA GAG CCG ACC ACC CAT CAA TAT GAT ACA GAT GCA GCA ACA ACC
    Tyr Ala Lys Glu Ile Glu Tyr Thr Gly Gln Phe Leu Ile Glu Pro Lys Pro Lys Glu Pro Thr Thr His Gln Tyr Asp Thr Asp Ala Ala Thr Thr

1066 ATT GCC TTT TTG AAG CAA TAT GGC TTA GAC AAT CAT TTT AAA TTA AAT CTA GAA GCC AAT CAT GCC ACA TTA GCC GGG CAT ACA TTC GAA CAT GAA
     Ile Ala Phe Leu Lys Gln Tyr Gly Leu Asp Asn His Phe Lys Leu Asn Leu Glu Ala Asn His Ala Thr Leu Ala Gly His Thr Phe Glu His Glu

1162 TTA CGC ATG GCA AGA GTA CAT GGT CTT CTT GGA TCT GTT GAT GCG AAC CAG GGT CAT CCT CTT TTA GGC TGG GAC ACG GAT GAA TTT CCC ACA GAT
     Leu Arg Met Ala Arg Val His Gly Leu Leu Gly Ser Val Asp Ala Asn Gln Gly His Pro Leu Leu Gly Trp Asp Thr Asp Glu Phe Pro Thr Asp

1258 TTA TAT TCT ACG ACA TTA GCA ATG TAC GAA ATC CTG CAA AAT GGC GGC CTT GGA AGC GGT GGC TTA AAC TTT GAC GCG AAG GTC AGA AGA TCT TCT
     Leu Tyr Ser Thr Thr Leu Ala Met Tyr Glu Ile Leu Gln Asn Gly Gly Leu Gly Ser Gly Gly Leu Asn Phe Asp Ala Lys Val Arg Arg Ser Ser

1354 TTT GAG CCT GAT GAT TTA GTA TAT GCC CAT ATT GCA GGG ATG GAT GCA TTT GCA AGA GGA TTG AAA GTA GCC CAC AAA TTA ATC GAA GAT CGT GTG
     Phe Glu Pro Asp Asp Leu Val Tyr Ala His Ile Ala Gly Met Asp Ala Phe Ala Arg Gly Leu Lys Val Ala His Lys Leu Ile Glu Asp Arg Val

1450 TTT GAA GAT GTG ATT CAA CAT CGT TAT CGC AGT TTT ACT GAA GGA ATT GGT CTT GAA ATT ACA GAA GGA AGA GCT AAT TTC CAT ACT CTT GAG CAA
     Phe Glu Asp Val Ile Gln His Arg Tyr Arg Ser Phe Thr Glu Gly Ile Gly Leu Glu Ile Thr Glu Gly Arg Ala Asn Phe His Thr Leu Glu Gln

1546 TAT GCG CTA AAT AAT AAA ACA ATT AAA AAT GAA TCT GGA AGA CAG GAG CGA TTA AAA CCT ATA TTG AAC CAA TAA
     Tyr Ala Leu Asn Asn Lys Thr Ile Lys Asn Glu Ser Gly Arg Gln Glu Arg Leu Lys Pro Ile Leu Asn Gln end

1621 CATTTTAGAA GTATAACAGG TGAAAGAAGA AAGCTACAGA TCCTGCTAGT AAGAAGAGAT AGCAGGATAG CCAAGTCAC ATAAACATCCC GTCATGATTC CATTACTTTT GCTTATGTTA

1741 TGACGGTAAT TTCTATAATT GGATTTATTT AGATGAATGT TTTTTTAAAA GGTTAAGGAG TTGAAAAAAT GAAGTATGT CATAGGAATAG ATCT
```

Figure 2 Sequence of the xylose isomerase gene with 5' and 3' flankingregions. 80 % of the nucleotide sequence have been determined on both strands. The amino acid sequence was deduced from the DNA sequence.The potential promoter hexamers are indicated with dots, the Shine-Dalgarno sequence is underlined. Underlinedamino acid sequences indicate regions homologous to the E. coli xylose isomerase.

The DNA sequence contains an open reading frame, starting with an ATG initiation codon, of 1320 nucleotides. A Shine-Dalgarno sequence (-14) AAGGAG, precedes the ATG initiation codon 8 nucleotides upstream. Six other ATG codons are present in the isomerase leader sequence but they are immediately followed by stop codons in all three reading frames. In the leader sequence we also recognized a long run of A's followed by two hexamers

TTGAAA and TAAGAT at positions -132 and -109 respectively. These structural features appear to be typical for *B. subtilis* promoter regions (10). The structural gene codes for an amino acid chain of 440 residues with a molecular weight of 49 680.

175 base pairs downstream from the TAA stop codon another AAGGAG Shine-Dalgarno sequence appears, 8 bp proximal to an ATG triplet most likely indicating the start of the xylulokinase coding region.

The codon usage for the xylose isomerase gene is summarized in Table I. No significant deviations can be observed when compared to the codon usage of the *B. subtilis* α-amylase gene (11) or the *B. pumilis* xylanase gene (12).

Table I: Codon usage for the *B. subtilis* xylose isomerase.

Phe	UUU	24	Ser	UCU	9	Tyr	UAU	15	Cys	UGU	1
	UUC	4		UCC	1		UAC	3		UGC	0
				UCA	1	Och	UAA	1	Umb	UGA	0
Leu	UUA	19		UCG	0	Amb	UAG	0	Trp	UGG	5
	UUG	7									
Leu	CUU	8	Pro	CCU	5	His	CAU	16	Arg	CGU	4
	CUC	0		CCC	2		CAC	2		CGC	2
	CUA	4		CCA	5	Gln	CAA	11		CGA	3
	CUG	1		CCG	1		CAG	3		CGG	0
Ile	AUU	11	Thr	ACU	6	Asn	AAU	16	Ser	AGU	3
	AUC	5		ACC	4		AAC	9		AGC	3
	AUA	1		ACA	15	Lys	AAA	20	Arg	AGA	10
				ACG	5		AAG	5		AGG	1
Met	AUG	13									
Val	GUU	5	Ala	GCU	8	Asp	GAU	23	Gly	GGU	4
	GUC	2		GCC	9		GAC	4		GGC	12
	GUA	9		GCA	20	Glu	GAA	21		GGA	12
	GUG	5		GCG	7		GAG	7		GGG	5

DISCUSSION

Within the sequenced DNA fragment of 1834 bp which can complement an *E. coli* xylose isomerase mutation there exists only one open reading frame of a length compatible with the xylose isomerase polypeptide. SDS-PAGE analysis of purified *B. subtilis* xylose isomerase (results not shown) revealed an apparent molecular weight of approx. 46 000 which is similar to the molecular weights reported for *E. coli* (2) and *Streptomyces albus* (13) xylose isomerases. This value is in agreement with the molecular weight of 49 680 deduced from the open reading frame. The molecular weight calculated from gel filtration experiments (unpublished observation) suggests that the native *B. subtilis* isomerase is a dimer. A dimeric structure is reported also for *Streptomyces albus* isomerase (13).

In addition to the *xyl*$^-$ complementation and the size agreement between the open reading frame and the xylose isomerase monomer, there is one more strong argument indicating that we are dealing with the structural gene for this enzyme. The amino acid sequence derived from the open reading frame shows more than 50% homology with the sequence reported for *E. coli* xylose isomerase (2,3). The total lengths of both sequences are almost identical and the homology extends from the beginning and almost till the end of the sequence published by Schellenberg *et al.* (2). The latter authors had confirmed the identification of the gene by an antibody assay in a hybrid selection and translation experiment. On the basis of the strong homology between the *Bacillus* and the *E. coli* enzymes (compare Fig.2) we can conclude that the shorter open reading frame published by Briggs *et al.* (3) must be due to a deviation in the nucleotide sequence leading to an artefactual stop codon. The above mentioned homology exceedes this stop codon and extends for an additional 100 amino acids which is in agreement with the data from Schellenberger *et al.* (2).

The strong homology between the xylose isomerases of a gram positive and a gram negative bacterium points to a high degree of conservation and infers that most of the primary structure is essential for its catalytic function.

The *B. subtilis* polypeptide harbors only one cysteine residue while the *E. coli* peptide contains four. The codon usage of the *B. subtilis* sequence does not show significant deviations when compared to other *Bacillus* genes nor does it show an obvious codon bias. Compared to *E. coli* *B. subtilis* prefers codons with a higher A/T content for several amino acids, reflecting the higher AT content of its DNA.

The absence of homology between the DNA sequences of both xylose isomerase genes each reflecting the overall base composition of the host organism, excludes the possibility of a recent horizontal exchange between both bacterial species.

ACKNOWLEDGEMENT

This work is supported by EC contract GBI-3-014-D and the Bundesministerium für Forschung und Technologie in Bonn.

REFERENCES

1. Wilhelm, M., Hollenberg, C.P. (1984) EMBO Journal 3, 2555-2560.
2. Schellenberg, G.D., Sarthy, A., Larson, A.E., Backer, M.P., Crabb, J.W., Lidstrom, M., Hall, B.D., Furlong, C.E. (1983) J.Biol.Chem. 259, 6826-6832.
3. Briggs, K.A., Lancashire, W.E., Hartley, B.S. (1984) EMBO Journal 3, 611-616.
4. Birnboim, H.C., Doly, J. (1979) Nucleic Acids Res. 7, 1513-1523.
5. Maniatis, T., Fritsch, E.F., Sambrook, J. (1982) in Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory Press, N.Y.
6. Herrmann, R.G., Whitfeld, P.R., Bottomley, W. (1980) Gene 8, 179-191.
7. Sanger, F., Nicklen, S., Coulson, A.P. (1977) Proc.Natl.Acad. Sci. USA 74, 5463-5467.
8. Vieira, J., Messing, J. (1982) Gene 19, 259-268.
9. Maxam, A.M., Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
10. Moran, C.P. Lang, N., LeGrice, St.F.U., Lee, G., Stephens, M., Sonenshein, A.L., Pero, J., Losick, R. (1982) Mol.Gen.Genet. 186, 339-346.
11. Yamazaki, H., Ohmura, K., Nakayama, A., Takeichi, Y., Otozai, K., Yamasaki, M., Tamura, G., Yamane, K. (1983) J. Bacteriol. 156, 327-337.
12. Fukusaki, E., Panbangred, W., Shinmyo, A., Okada, H. (1984) FEBS Letters 171, 197-201.
13. Hogue-Angeletti, R.A. (1975) J.Biol.Chem. 250, 7814.