FEBS Letters 433 (1998) 108–112 FEBS 20650

Evolution of new protein function: recombinational enhancer Fis originated by horizontal gene transfer from the transcriptional regulator NtrC

Enrique Morett[a,b,*], Peer Bork[b,c]

[a] *Instituto de Biotecnología, Universidad Nacional Autónoma de México, AP 510-3 Cuernavaca, Mor., Mexico*
[b] *European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany*
[c] *Max-Delbrück-Centrum für Molekulare Medizin, Robert-Roessle Str. 10, 13122 Berlin-Buch, Germany*

Received 5 May 1998; revised version received 14 July 1998

Abstract New protein function is thought to evolve mostly by gene duplication and divergence. Here we present phylogenetic evidence that the multifunctional protein Fis of the γ proteobacterial species derived from the COOH-terminal domain of an ancestral α proteobacterial NtrC transcriptional regulatory protein. All of the known enterobacterial *fis* genes are preceded by an open reading frame, named *yhdG*, that is highly similar to *nifR3*, a gene that forms an operon with *ntrC* in several α proteobacterial species. Thus, we propose that *yhdG* and *fis* were acquired by a lineage ancestral to the γ proteobacteria in a single horizontal gene transfer event, and later diverged to their present functions.

Key words: Protein evolution; Phylogenetic analysis; Enhancer-binding protein

1. Introduction

The lateral transfer of genes is a major force driving the evolution of microorganisms, facilitating dissemination of genes and allowing for the evolution of new functions to occur. There is ample evidence that bacterial chromosomes contain DNA regions inherited by horizontal gene transfer [1,2]. Several attempts have been made to quantify the amount of horizontally transferred genes into *Escherichia coli*. Estimates range from 6% to 16% ([3], reviewed in [4]). Recently, Lawrence and Ochman [4], using codon usage patterns, calculated that at least 15% of the genes of this bacteria have been inherited in the recent past only from species departing from the GC content and the typical codon usage of *E. coli*. The same authors estimated that at least 10% of the *Salmonella enterica* serotype Thyphimurium genes arose through horizontal transfer [5]. In spite of this large proportion the origin of the laterally inherited genes has been difficult to trace back [1,6,7].

Fis (factor for inversion stimulation) is a multifunctional protein involved in many site-specific recombination events, regulation of gene expression and *oriC*-directed initiation of chromosomal replication (reviewed in [8]). This protein was first identified in *E. coli* as a cellular factor required to stimulate DNA inversions catalyzed by the site-specific recombinases Hin, which controls flagellar phase variations in *S. enterica* serotype Thyphimurium [9] and Gin, which controls tail fiber expression in bacteriophage Mu [10]. Fis also stimulates lambda phage integration and excision [11] and transposition by Tn*5* and IS*50* [12]. Additionally, by binding to *oriC*, Fis influences replication of the *E. coli* chromosome [13,14].

All the above processes involve higher-order nucleoprotein complexes. Fis is a 98 amino acids long, α helical, basic, dimeric protein that binds to highly degenerated DNA sites [8,15,16]. Upon binding Fis induces a bend in the DNA [8]. However, not all the known functions of Fis are related to specialized DNA recombination and DNA replication. Fis is also a transcriptional activator of components of the translational machinery, including rRNA and tRNA genes. Recently, González-Gil et al. [17] reported that Fis is also involved in sugar and nucleotide metabolism in *E. coli*. Osuna et al. [18] have proposed that these functions may offer competitive advantage to the cell.

In spite of the central role that Fis plays in the expression of ribosomal and transfer RNA, to date *fis* has only been identified in a few species belonging to the γ proteobacteria branch of the Eubacteria. *fis* forms an operon with *yhdG*, a gene of unknown function, and they are not present in any of the completely sequenced bacterial genomes other than those of the γ proteobacterial *E. coli* and *Haemophilus influenzae* species.

Sequence analysis revealed that Fis has significant similarity to NtrC, a protein that belongs to the enhancer-binding protein (EBP) family of transcriptional regulatory proteins [19,20]. Here we present phylogenetic evidence showing that Fis originated from the COOH-terminal domain of an ancestral α proteobacterial NtrC and acquired by lateral transfer, together with *yhdG*, by a lineage ancestral to the *Pasteurelaceae* and *Enterobacteriaceae*.

2. Materials and methods

2.1. Sequence analysis

A multiple amino acid sequence alignment of the known Fis and the carboxy-terminal domain of several NtrC proteins was produced using the multiple sequence alignment program ClustalW [34]. Proteins from *S. enterica* and *Klebsiella pneumoniae* were not included because they are 100% identical to the *E. coli* protein. The alignment was manually refined at positions were clear misalignments were produced by the algorithm. When annotated, protein sequences are labeled by their SwissProt name. Other proteins are: *Rhodospirillum rubrum* NtrC, named here NTRC_RHORU (EMBL: G927310), *Agrobacterium tumefaciens* NtrC (NTRC_AGRTU; EMBL: G142244), *Rhodobacter sphaeroides* NtrC (NTRC_RHOSP; EMBL: G468815), *Thiobacillus ferrooxidans* NtrC (NTRC_THIFE; EMBL: G310877), *Erwinia carotovora* Fis (FIS_ERWCA; EMBL: G2773326)

*Corresponding author. Departamento de Reconocimiento Molecular y Bioestructura, Instituto de Biotecnología, Universidad Nacional Autónoma de México, AP. 510-3, Cuernavaca, Mor. Mexico.
E-mail: emorett@ibt.unam.mx

PII: S0014-5793(98)00888-6

```
             1                                                                                                           114
NTRC_AGRTU lyp.Qevitr.eii.eqelqsdvpdspl.dkmavrt.gsltisqaveenmrdyfasfgdglpppgLydrvlrElEypLilaaltaTrGNqikAAdlLGlnRnTLrkKirelgvs
NTRC_RHIME lyp.Qdvitr.eii.enelrseipdspi.ekaaars.gslsisqaveenmrqyfasfgdalppsgLydrvlaEmEypLilaaltaTrGNqikAAdlLGlnRnTLrkKirelgvs
NTRC_BRASR lyp.Qdvita.svi.dgela....ppavtsgstatv.gvdnlggaveayLsshfsgfpngvpppgLyhrilkEiEipLltaalaaTrGNqirAAdlLGlnRnTLrkKirdldiq
NTRC_RHORU lys.Qevigv.evi.eaels.s.sappildeg.de..saeglstkverhLrhyfeahgealppagLherisrEvErpLislcleaTrGNqikAAqlLGlnRnTLrkKirdldiq
NTRC_AZOBR lys.Qevigl.dvv.eaela.d.ttpaa.q.pveep.qgeglsaaverhLkdyfaahkdgmpsngLydrvlrEvErpLislslsaTrGNqikAAqlLGlnRnTLrkKirdldiq
NTRC_RHOSP tsa.eaeitraev..eavlg.n..qpam.e.plkgggegeklsssvarhLrryfdlhggalpppgvyqrilrEvEppLieialdaTaGNqakcAdlLGinRnTLrkKitdldir
NTRC_RHOCA tas.eeeitrtev..efvlg.n..qpav.e.plgaggegeklsasiarhLrryfdlhggnlpppgLydrilaEmEapLieialdaTgGNqakcAdlLGinRnTLrkKitdldiq
  FIS_ECOLI mfe.Qrvns..dvltvstv..n.sqdqvtqkp........lrdsvkqaLknyfa.qlngqdvndLyelvlaEvEqpLldmvmqyTrGNqtrAAlmmGinRgTLrkKlkkygmn
  FIS_SERMA mfe.Qrvns..dvltvstv..n.sqdqvtqkp........lrdsvkqaLknyfa.qlngqdvndLyelvlaEvEqpLldmvmqyTrGNqtrAAlmmGinRgTLrkKlkkygmn
  FIS_PROVU mfe.Qrvns..dvltvatv..n.sqdqvtqkp........lrdsvkqaLknyfa.qlngqdvndLyelvlaEvEqpLldmvmqyTrGNqtrAAqmmGinRgTLrkKlkkygmn
  FIS_ERWCA mfe.Qrvns..dvltvstv..n.sqaqvtqkp........lrdsvkqaLknyfa.qlngqdvsdLyelvlaEvEqpLldmvmqyTrGNqtrAAlmmGinRgTLrkKlkkygmn
  FIS_HAEIN mle.Qqrns..daltvsvl..n.aqsqvtskp........lrdsvkqaLrnyla.qldgqdvndLyelvlaEvEhpmldmimqyTrGNqtrAAnmLGinRgTLrkKlkkygmg
  FIS_PASHA mle.Qq.pt..npltvtml..n.aqaqqvnkp........lrdnvkaaLknyls.qlngedpteLyelvlsEiEhpmldmvmqyTrGNqtrAAtmLGinRgTLrkKlkkygmg
 NTRC_SALTY maagQevliq.dlp..gelf.e.astpdspsh.l......ppdswatlLaqwadralrsghqn.LlseaqpElErtLlttalrhTqGhkqeAArlLGwgRnTLtrKlkelgme
 NTRC_ECOLI maagQevliq.dlp..gelf.e.stvaestsq.m......qpdswatlLaqwadralrsghqn.LlseaqpElErtLlttalrhTqGhkqeAArlLGwgRnTLtrKlkelgme
 NTRC_KLEPN maagQevltq.dlp..self.e.taipdnptq.m......lpdswatlLgqwadralrsghqn.LlseaqpEmErtLlttalrhTqGhkqeAArlLGwgRnTLtrKlkelgme
 NTRC_PROVU mtasQeimpq.dlp..seir.l.adekakninrl......tsqhwsqhLslwadealgegken.ilndalpqfErtLllsalayTqGhkqdAArlLGwgRnTLtrKlkelgme
 NTRC_THIFE maptrevqvt.dlp..pema.a.spmsammpd.p......ntqdwrllLgavvrh.vsrreea.LldkiqplfErcLleaalqhTrGhkqsAAhkLGwgRnTLtrKlkelgma
 ACOR_CLOMA lsk.ndmisy..l...sip.....eyine.................needtisttnfnstrp....dk.leEtEksLivkaleysgGNvvkAskliGigksTLyrKikkyels
 FLBD_CAUCR lsa.gpeiee......fair.....lpdgqpm.........aapdvavargaqmaa..daasra.fvgstvaEvEqqLiidtlehclGNrthAAniLGisirTLrnKlkeysda
```

Fig. 1. Multiple sequence alignment of Fis and the COOH-terminal domain of NtrC. The positions exclusively conserved between Fis and the α or the γ proteobacterial NtrC proteins are in bold. The α proteobacterial NtrC protein names are in italics. Positions conserved in more than 75% of the sequences are in uppercase. When available, SWISS-PROT codes were used to identify the proteins (see Section 2 for the sequence data bank entry and the definition of the sequence names for the rest of the protein sequences). The multiple sequence alignment was generated using the multiple sequence analysis package CLUSTALW [34] and manually refined. The carboxy-terminal domain sequences of the enhancer-binding proteins AcoR of *Clostridium magnum* and FlbD of *Caulobacter crescentus*, that were used to root the phylogenetic tree of Fig. 2, are also shown.

Proteus vulgaris Fis (FIS_PROVU; EMBL: G2773316), *Pasteurella haemolytica* FIS (FIS_PASHA; EMBL: G2731435), and *Serratia marcescens* FIS (FIS_SERMA; EMBL: G2773310).

The proteins with the greatest similarity to YhdG and NifR3 were obtained by PSI_BLAST searches [29]. A multiple amino acid sequence alignment of 29 proteins was generated with ClustalW and manually refined at positions were ambiguity was observed. Proteins not annotated in SwissProt are: *K. pneumoniae* YhdG, named here YHDG_KLEPN (EMBL: AF040380), *S. marcescens* YhdG (YHDG_SERMA; EMBL: AF040378), *E. carotovora* YhdG (YHDG_ERWCA; EMBL: AF040381), *P. vulgaris* YhdG (YHDG_PROVU; EMBL: AF040379), *Borrelia burgdorferi* YhdG (YHDG_BORBU; Genebank: AE001133), *Mycobacterium tuberculosis* YntB (YNTB_MYCTU; EMBL: G2916881), and *Ralstonia eutropha* ORF2 (ORF2_RALEU; EMBL: G3046395).

2.2. Phylogenetic analyses

Parsimony- and distance-based phylogenies were reconstructed using the J. Felsenstein's PHYLIP package, version 3.5, for UNIX (J. Felsenstein, 1993, Department of Genetics, University of Washington, Seattle, USA, distributed by the author). The Fis/NtrC and YhdG/NifR3 multiple sequence alignments were subjected to bootstrapping resampling (100 times), using the program SEQBOOT. Parsimony-based phylogenies of the resampled alignments were estimated using the PROTPARS program. Genetic distances of the resampled alignments were calculated with the program PROTDIST, and trees constructed with the FITCH program. Consensus trees were calculated using the program CONSENSE. For tree representation and edition the program TreeView 1.5 (Page, R., University of Glasgow, UK) was used on a Macintosh computer.

3. Results and discussion

3.1. Phylogenetic relationship of Fis and NtrC

A multiple sequence alignment of the known Fis and the COOH-terminal domain of NtrC proteins is shown in Fig. 1. NtrC belongs to the large EBP family of regulatory proteins that activate transcription from promoters recognized by the σ^{54}-holoenzyme form of the RNA polymerase. These proteins are generally composed of three structurally and functionally different domains. The NH_2-terminal domain of NtrC is a response regulator domain of the two component family of the sensor-transduction regulatory proteins. The central domain is common to the whole EBP family and has the transcriptional activation functions, whereas the COOH-terminal

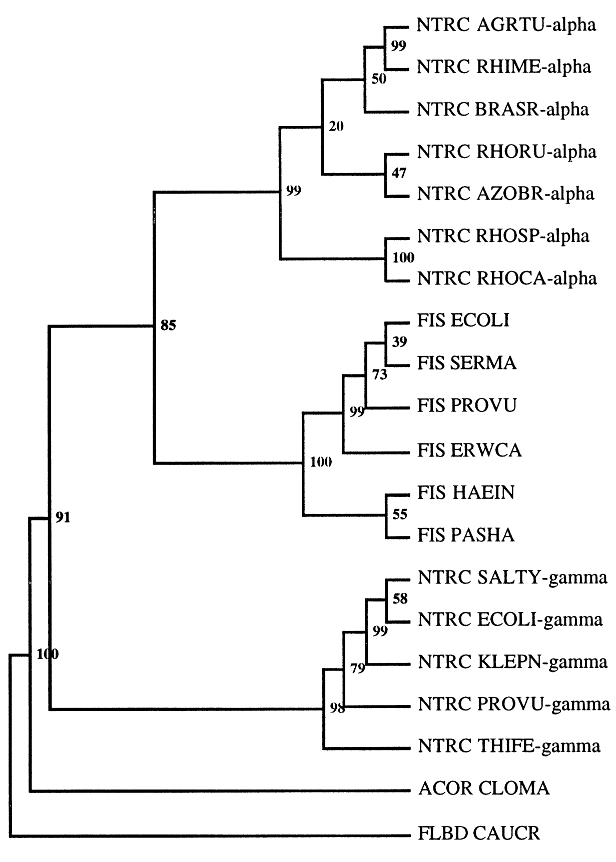

Fig. 2. Consensus phylogenetic tree of Fis and the COOH-terminal domain of NtrC. The multiple sequence alignment of Fig. 1 was resampled 100 times using the bootstrapping program SEQBOOT of the PHYLIP package (see Section 2), and the most parsimonious trees for each sample obtained using the program PROTPARS [34]. The consensus tree shown was generated using the program CONSENSE of the same package. The carboxy-terminal domain sequences of the enhancer-binding proteins AcoR of *Clostridium magnum* and FlbD of *Caulobacter crescentus* were used to root the tree. The Fis proteins clustered with the α proteobacterial (marked alpha) subset of the NtrC proteins apart from the γ proteobacterial (marked gamma) NtrC proteins.

```
PYRD_ENTFA qaieaggadgfsmintllgmridlktrkpilan..gtgglsgpaIkpvairlirqv.as.vsqlpiigmGgvqtvddvlemfm.....aGasavgvGtanftdPyicpk
PYRD_BACSU laieeagadgltmintligmridlktgkpilan..ktgglsgpavkpvairmvyev.sq.mvnipiigmGgvqtaedalefll.....aGasavavGtanfvnPfacpe
PYRD_METJA qavvdagvdglvaintvrgmaidiRakkpilan..kfgglsgkaIksigikvvwdl.ye.nfdvpiigvGgImsgedaieymm.....aGasavqiGsgvyyrgydifk
PYDA_LACLC .qfpltyvnsvnsig..nglfidpeaesvvikpkdgfggiggayIkptalanvrafytrlkpeiqiigtGgIetgqdafehll.....cGatmlqiGtalhkegpaifd
PYRD_YEAST .efplayvnsinsig..nglfidvekesvvvkpkngfggiggeyvkptalanvrafytrlrpeikvigtGgIksgkdafehll.....cGasmlqiGtelqkegvkife
SMM1_YEAST ...qlvkrlcatGi...tnltvHcRkte..mr..nreqpit.dyIaeiyeic.......qannvslivNGaIrdrsh.fhdlqanhwkntniggmiaecaerd.ptvfd

YHDG_PROVU .mrigq.....yqlkn.....cliaA.PMagvtdrpfRsl.c.ydmga.........................gmtv.seMlssnpqvwq.tdksrlr..mvhsde...
YHDG_HAEIN .mrigs.....yqlrn.....rvllA.PMagitdqpfRrl.c.ayyga.........................gltf.seMmstnpqvwh.teksklr..lahsed...
NIR3_RHILP .ms..........vrn.....rvvlA.PMsgvtdmpfRel.a.wrfga.........................glvv.teMvas.relvndtaeswsr.lkaagfr...
NIR3_RHOCA .mtisld.....slrl...dppvllA.PMagitdlpfRrm.v.arfga.........................glvv.seMvas.geml..takpsvr.akaqaeltag
NIR3_AZOBR .maiqigt...isle.....gpvilA.PMsgvtdlpfRrl.v.kqsgc.........................glvv.seMvas.qamirenrqt.lr..mvecepeqf
YNTB_MYCLE ........................mlA.PMagvtnvtfRtl.c.refgstrgnllrpapphlrspleqsqvgtisglyv.ceMvta.ralvernavtihm.ttfapdes..
YNTB_MYCTU alrigp.....ielas.....pvvlA.PMagvtnvafRal.c.rq.....................leqskvgtvsglyv.ceMvta.ralierhpvtmhm.ttfsades..
YOHI_ECOLI .....................mrvllA.PMegvldslvRel.ltevndy.........................dlci.tefvrvvdqllp.vkvfhri..cpelqnasr
YOHI_HAEIN .....................mrvilA.PMqgvldpfvRel.ltevndy.........................dlci.tefvrvvdqllp.ekvfyrl..cpelknqgf
ORF2_RALEU .....................msrlflA.PMegladyvlRdv.ltdtggy.........................dgcv.sefvrvtgsllp.arvyere..tpeilaggy
YJBN_HAEIN .mpqnlp....hfyrg.....rfsvA.PMldwttrhcRyf.hrqfskh.........................ally.teMvta.paiihakydhld....fdlqen..
YJBN_ECOLI .mpektd....vhwsg.....rfsvA.PMldwtdrhcRyf.lrllsrn.........................tlly.teMvtt.gaiihgkgdyla....yseeeh..
Y926_SYNY3 .mtkifa....dssin.....plsvA.PMmdhtdrhfRyf.lrqltrh.........................tlly.teMita.qailhgdrqrll...nfspeek..
YJBN_BORBU .............mlmn....rkisiA.PMvnitdehfRyl.irllskk.........................vtly.tpMisa.ksiimgdvkkivk...qnples..
YL05_YEAST liteetdplhiiktrqkthgrpvtiAgPMvrysklpfRql.c.reynv.........................divy.spMila.reyvrneharis...dlstnnedt
YMI0_YEAST lmqkltgrqlfdkigr...ptriv.A.PMvdqselawR.i.lsrryga.........................tlay.tpMlha..klfatskkyred.nwssldgssv
YQI2_CAEEL .............men.....dpsf..tllpwla..fRql.v.rvydv.........................dvcf.tpMiya.knfiesekcrss...elsvcegds
PYDB_LACLC rlsvklp...gldlkn.....piipAsgcfgfgeeyakyydlnkl.gsimv..kattlhprfgnptpr......vaetasgMlna.iglqnpglevimteklpwlnenf.
PYRD_ENTFA plavsip...gltlkn.....piipAsgcfgfgeeyanyydldql.gsimi..kattpqarygnptpr......vaetpsgMlna.iglqnpgldvvmqeklpklek.y.
PYRD_BACSU mlevklp...gldlkn.....piipAsgcfgfgkefsRfydlscl.gaimi..kattkeprfgnptpr......vaetgagMlna.iglqnpgldsvlhhelpwleqfdt
PYRD_METJA mlktnic...giefkn.....pvflAsgimgetgsalk..riakg.gagavttksiglnpnpghknpt......ivevyggflna.mglpnpgvdeyle.eiekvrdeln
PYDA_LACLC mlnttfa...nakfan.....pfmnAsgvhcmtiedleelkasqa.gayit..ksstlekregnplpr......yvdlelgsins.mglpnlgfdyyld.yvlknqkena
PYRD_YEAST slttkfl...nntyen.....pfmnAsgvhcmttqeldelanska.gafit..ksattleregnpepr......yisvplgsins.mglpnegidyyls.yvlnrqkny.
SMM1_YEAST mvtyag..............klvlA.PMvragelptR.l.malahgadlv..wspeiidkkliqcvr........kentalqt.......vdyvvp.skvqtrpetl

          111                                                                                                  220
YHDG_ECOLI .....Pgi....rtvQiaGsdpkemadaArinvesg.aqiIdiNmGCPa.kkVnr.klaGsaLlqypdvvksiltevvnav....dvPVtlKiRtg.......wapehrn
YHDG_KLEPN .....Pgi....rtvQiaGsvpkemaaaArinvesg.aqiIdiNmGCPa.kkVnr.klaGsaLlqypdqvksiltavvkav....dvPVtlKiRtg.......wepehrn
YHDG_SALTY .....Pgi....rtvQiaGsdpvemadaArinvesg.aqiIdiNmGCPa.kkVnr.klaGsaLlqypdlvksiligvvnav....dvPVtlKiRtg.......wapehrn
YHDG_SERMA .....Pgi....rsvQiaGcdpddmaaaArisvasg.aqlIdiNmGCPa.kkVnr.klaGsaLlqypdlvkrilhavvdav....dvPVtlKiRtg.......wapehrn
YHDG_ERWCA .....Pgi....ravQiaGcdpdemaaarrinadsg.aqiIdiNmGCPa.kkVnr.kmaGsaLlqypdlvkqilstvvkav....dvPVtlKiRtg.......wapehrn
YHDG_PROVU .....lgi....rsvQiaGsdpvdmaaaAkinadsg.aqiIdiNmGCPa.kkVnk.klaGsaLlrhpdlvaeilsavvnav....dvPVtlKiRtg.......wspdern
YHDG_HAEIN .....lgl....navQiaGsdplemaqaAainveyg.aqiIdiNmGCPa.kkVnr.klaGsaLlqfpdlvekilrevvsav....nvPVtlKiRtg.......wdksnrn
NIR3_RHILP .....P......hmvQlaGreahwmaeaAkiaadhg.adiIdiNmGCPa.kkVig.gysGsaLmrdpdhalglieatvkav....diPVtlKmRlg.......wdensin
NIR3_RHOCA l....P......tsvQlaGreaapmaeaAkivadmg.aeiIdiNmGCPa.kkVtg.glsGaaLmrnpdhalrlieavvgav....dlPVtlKmRlg.......wdedqln
NIR3_AZOBR .....P......mavQlaGcepdvmaeaAklnedrg.aaiIdiNfGCPv.kkVvn.ghaGssLmrdealaarileatakav....tiPVtlKmRkg.......wddssln
YNTB_MYCLE .....P......rslQlytvdpattytaAkmvadegladhIdmNfGCPm.pkVtr.rggGaaLpykrrlfgqivaaavrate.gtkiPVtvKfRig.......iddehht
YNTB_MYCTU .....P......rslQlytvdpdttyaaArmiagegladhIdmNfGCPv.pkVtk.rggGaaLpfkrrlfgqivaaavrate.gtdiPVtvKfRig.......iddahht
YOHI_ECOLI t....Psgtl..vrvQllGqfpqwlaenAaravelg.swgvdlNcGCPs.ktVng.sggGatLlkdpeliyqgakamreavp..ahlPVsvKvRlg.......wdsgekk
YOHI_HAEIN tssgtP......vrvQllGghpkclaenAiraidlg.shgIdlNcGCPs.ktVng.sngGaaLlkpeliyratqalrravp..sefPVsvKvRlg.......wddisq.
ORF2_RALEU trsgtP......mviQllGsdpewlarnAayaatls.phgIdlNfGCPa.kvVnr.hggGamLltnpellnrivasvraavp..ahiaVtaKmRlg.......vsdasl.
YJBN_HAEIN .....P......valQlgGsdpeqlkycAklaeerg.yheInlNvGCPs.drVqn.gmfGacLmakadlvadcveqmqsav....kiPVtvKtRig......ideldsye
YJBN_ECOLI .....P......valQlgGsdpaalaqcAklaearg.ydeInlNvGCPs.drVqn.gmfGacLmgnaqlvadcvkamrdvv....siPVtvKtRig......iddqdsye
Y926_SYNY3 .....P......valQlgGddpqllaecArigqdwg.ydeInlNvGCPs.drVqs.gnfGacLmaqpdlvaqcvsamqkav....eiPVtvKhRig......idhrdsye
YJBN_BORBU .....P......iaiQiatnskndalkaiqilekhfnfdeynlNvGCPs.lkiqn.ancGacLmgnanqvgeilktmkent....nkPisiKhRlgirplssdyknesys
YL05_YEAST .....P......livQvgvnnvadllkfvemvapyc..dgIgiNcGCPikeqir..egiGcaLiynsdllcsmvhavkdkyg..dklrietKiRih........ealde.
YMI0_YEAST dr...P......lvvQfcandpeyllaaAklvedkc..davdlNlGCP..qgiakkghyGsfLmeewdlihnlintlhknl....kvPVtaKiRi........fddcek.
YQI2_CAEEL .....P......livQfatddpfvlseaAemvykcs..tgvdlNcGCP.khdVrs.kgfGsaLlskpelladmvrqtrarip.dpdfsVslKiRin........hdiek.
PYDB_LACLC .....Pelp...iianvaGseeadyvavcakigdranvkaIelNisCP...nVkh...gGqafgtdpevaaalvkac.kav...skvPlyvKlspn......vtdivpia
PYRD_ENTFA .....Pnlp...iianvaGaceedyvavcakigqapnvkaIelNisCP...nVkh...gGiafgtdpevafqltqavkkva....svPiyvKlspn......vtdivpia
PYRD_BACSU .....P......iianvaGsqvddyvevAehiskapnvhalelNisCP...nVkt...gGiafgtnpemaadltkavkevs....dvPVyvKlspn......vaniteia
PYRD_METJA r....mdv...riigsiyGkdeeefaevAkkmer..yvdiIelNisCP...hak...gyGatigqnpdlsydvckavkkav....kiPVfaKltpn......vtdiieia
PYDA_LACLC qeg..P......iffsiaGmsaaeniamlkkiqesdfsgitelNlsCP...nVpgkpqlaydfeateklkevftfftkplg..vklPpyfdlvhf....dimaeiln..
PYRD_YEAST .....Pdap..aiffsvaGmsidenlnllrkiqdsefngitelNlsCP...nVpgkpqvaydfdltketlekvfaffkkplg..vklPpyfdfahf....dimakiln..
SMM1_YEAST vfrtyPklessklifQigsaspalatqaAlkvind..vsgIdiNaGCP...khfsihsgmGsaLlrtpdtlclilkelvknvgnphskPisvKiRll......dtkqdtl.

          221                                                                                                  329
YHDG_ECOLI .ceeiaqlaedcGi...qaltiHgRtra.clf....ngeaeydsIravkqkv.........sipviaNGdItdplkaravld.y...tGadalmiGraaqgrPwi.fr
YHDG_KLEPN .cveiaqlaeecGi...qaltiHgRtra.clf....ngdaeydsIravkqkv.........sipviaNGdItdplkaravld.y...tGadalmiGraaqgrPwi.fr
YHDG_SALTY .cveiaqlaedcGi...qaltiHgRtra.clf....ngeaeydsIravkqkv.........sipiiaNGdItnphkaravld.y...tGadalmiGraaqgrPwi.fr
YHDG_SERMA .cveiaqlaedcGi...qaltiHgRtra.clf....ngdaeydsIravkqsv.........sipiiaNGdItdphkaravld.y...tGadalmiGraaqgrPwi.fr
YHDG_ERWCA .cveiaklaedcGi...qaltiHgRtra.clf....ngfaeydsIravkqav.........sipiiaNGdItdphkaravld.y...tGadalmiGraaqgrPwi.fr
YHDG_PROVU .cveiaklaercGi...qaltiHgRtra.clf....kgeaeydsIrtvkqnv.........sipiiadGdItdplkaravld.y...tGadalmiGalaqgrPwi.fr
YHDG_HAEIN .cvqigkiaeqcGi...qaltvHgRtra.clf....egeaeydnIkavkqai.........aipviaNGdIdsarkakfvln.y...tGadaimiGraalgnPwl.fq
NIR3_RHILP .apdiarraeasGi...qlvtiHgRtrm.qfy....egradwdaIravrevi.........svpliaNGdvetahdaqeilr.r...sGadavmiGrgcqgrPwh.ag
NIR3_RHOCA .aaeiaaraeaaGv...kmiviHgRtrm.qfy....tgaadwraIaavraav.........svpvvaNGdItdaasarrald.q...sGaagvmvGrgaqgaPwr.la
NIR3_AZOBR .aprlariaeecGi...klvtvHgRtre.qfy....ngtadwsfIrsvkeav.........sipvvvNGdItsfdavdrala.e...sGadgvmiGrgaygrPwf.pa
YNTB_MYCLE .hldagriaeaeGa...aavalHaRtaa.qry....sgtadwaqIaqlkqqv.........rtipvlgNGdIydasdalamma.v...tGcdgvviGrgclgrPwl.fa
YNTB_MYCTU .hldagriaeaeGa...aavalHaRtaa.qry....sgtadweqIarlkqhv.........rtipvlgNGdIydagdalamms.t...tGcdgvviGrgclgrPwl.fa
YOHI_ECOLI ..feiadavqqaGa...telvvHgRtke.qgyr...aehidwqaIgdirqrl.........nipviaNGeIwdwqsaqqcma.i...sGcdavmiGrgalniPnl.sr
YOHI_HAEIN .afeiadaveqgGa...teitvHgRtka.dgyr...adrinwkkIsevrerl.........sipviaNGeIwhwqdgqdcls.q...tGcqdlmvGrgalniPnl.sh
ORF2_RALEU .aidcatalaegGa...aslvvHaRtrd.hgyr....ppahwdwIariaaav.........dvpviaNGdvwtvadwercra.v...sGcadvmiGrgavsdPflalr
YJBN_HAEIN flcdfiekvqgkGc...qefiiHaR.ka.wls...glspkenreIppldyirvyql.krdfphltiaiNGgIktieemkqhl.....qyvdgvmvGreayqnPsl.lg
YJBN_ECOLI flcdfintvsgkGec..emfiiHaR.ka.wls...glspkenreIppldyprvyql.krdfphltmsiNGgIksleeakahl.....qhmdgvmvGreayqnPgi.la
Y926_SYNY3 dlvhfveivanaGc...qrftvHaR.ka.wlq...glspkenrtIpplryedvyql.kkdfpqllieiNGgItqteqiqqhl.....shvdavmvGraayenPyl.fa
YJBN_BORBU evknfvkiisdfGi...knfivHaRi.a.ilk...gfspknnrnIpklrhefvykl.kkenknlfieiNGgInnseqikkhl.....sfvdsvmiGrsafenPyf.ia
YL05_YEAST .tvelcrklcdaGv...dwitiHgRtrr..tr...ssqpanldaIkyiieni.s.....dknvpvaNGdcfklsdleritk.y...tGahgvmavrgllsnPal.fa
YMI0_YEAST .slnyakmvldaGa...qfltvHgRvre.q.kg.qktglanwetIkylrdnl........pketvffaNGnIlypedisrcme.h...iGadavmsaegnlynPgv.fn
YQI2_CAEEL .tvdlcrkaeaaGv...thltvHgRtps...q...raepidiqalrivkdsv.........svpiiaNGgIttreealflae.q...tGvdgimaanglldnPal.fa
PYDB_LACLC kaveaagadgltmintlmgvrfdlktrqpilan..itgglsgpaIkpvalklihgv.ag.dvdipiigmGgvanagqvlemvm.....aGasrvavGtanfadPfvcpk
```

Fig. 3. Multiple sequence alignment of NifR3, YhdG and the proteins that showed the highest similarity to them in PSI-BLAST searches [29]. Positions conserved in more than 75% of the sequences are in uppercase. When available, SWISS-PROT codes were used to identify the proteins (see Section 2 for the sequence data bank entry and the definition of the sequence names for the rest of the protein sequences). The multiple sequence alignment was generated using the multiple sequence analysis package CLUSTALW [34] and manually refined.

domain, of about one hundred amino acids long, has a specific DNA-binding function (reviewed in [21]).

From the alignment it is evident that Fis shares significant sequence identity with the COOH-terminal domain of NtrC, but particularly with the α proteobacterial proteins (the names of these proteins are in italics in Fig. 1). Of the 98 amino acids of Fis, 16 are exclusively in common with this subset of the NtrC proteins. In contrast only seven positions are uniquely shared with the γ proteobacterial NtrC proteins. On average the Fis proteins are 40% identical to the α proteobacterial sequences, contrasting to 27% with the γ proteobacterial NtrC proteins. In the reported structure of Fis the NH_2-terminal region was not well resolved, apparently because the first 26 amino acids were disordered in the crystals [15]. Interestingly, this region of Fis showed the least similarity to NtrC. If this part of the protein is not taken into account the average identity of Fis with the COOH-terminal domain of the α and γ proteobacterial NtrC proteins increases to 50% and 33%, respectively.

To estimate the evolutionary relationship of the Fis and the NtrC proteins we carried out parsimony- and distance-based phylogenetic analyses. A parsimony-based tree is shown in Fig. 2. As expected from the alignment, Fis and the α proteobacterial subset of the NtrC proteins clustered together and apart from the γ proteobacterial NtrC proteins. The *Clostridium magnum* AcoR and the *Caulobacter crescentus* FlbD proteins were used to root the tree. A similar tree topology was obtained by genetic distance analysis (data not shown).

We have previously analyzed the evolutionary relationships of the EBPs [21]. From these analyses we concluded that the NtrC proteins are a monophyletic group, and the trees produced by these proteins agreed with the organismal phylogeny. Thus, it is clear that Fis originated from an ancestral α proteobacterial NtrC protein and it was acquired by a lineage ancestral to the *Enterobacteriaceae* and *Pasteurellaceae* γ proteobacteria by horizontal gene transfer.

3.2. Phylogenetic relationship of YhdG and NifR3

In all the bacteria in which Fis has been reported (except for *P. haemolytica*, where the sequence upstream of *fis* is not known) it forms part of an operon with an Orf of unknown function, denominated *yhdG* in *E. coli* [18,22–24]. YhdG belongs to family 34 of uncharacterized proteins [25]. Sequence analysis suggests that it has a TIM barrel structure with typical phosphate-binding site and it is likely to be an oxydoreductase, distantly related to PyrD, Smm1 and His4 ([25]; see also Fig. 3). *yhdG* is highly similar to *NifR3*, a gene located upstream of and cotranscribed with *ntrC* in *Rhodobacter capsulatus* [26], *Rhizobium etli* [27], and *Azospirillum brasilence* [28] (Fig. 3). Fig. 4 shows a parsimony-based tree of the known γ proteobacterial YhdG proteins and the proteins that showed the greatest sequence similarity to them in PSI-BLAST searches [29]. It is remarkable that the NifR3 proteins are the closest relatives to YhdG (Fig. 4). Thus, it is likely that *yhdG* and *fis* were acquired by an ancestor of the γ proteobacterial species by a single gene transfer in event (see below).

3.3. Amelioration of fis and yhdG sequences

Recent horizontally transferred genes can readily be detected by their unusual GC content and atypical codon usage. However, genes acquired in the distant past gradually adjust to the host GC and codon usage by mutational biases in a process termed amelioration [4,5]. Nucleotide sequence analysis of *yhdG* and *fis* of *E. coli* did not reveal any significant codon position GC content variation with the average chromosome, suggesting that these genes have completely ameliorated since their introgression into the γ proteobacterial species (data not shown). The fact that *yhdG* and *fis* genes are present in *H. influenzae*, a member of the *Pasteurellaceae* branch of the γ proteobacteria, indicates that the gene transfer event occurred before the split of this taxon from the *Enterobacteriaceae*. Assuming a divergence time of about 100–160 Mya for the closely related *E. coli* and *S. enterica* species [30], the split of *Pasteurellaceae* from *Enterobacteriaceae* must have occurred long before. This result is in agreement with Ochman and Lawrence's [5] calculation that an horizontally transferred sequence will completely ameliorate in a maximum 400 My.

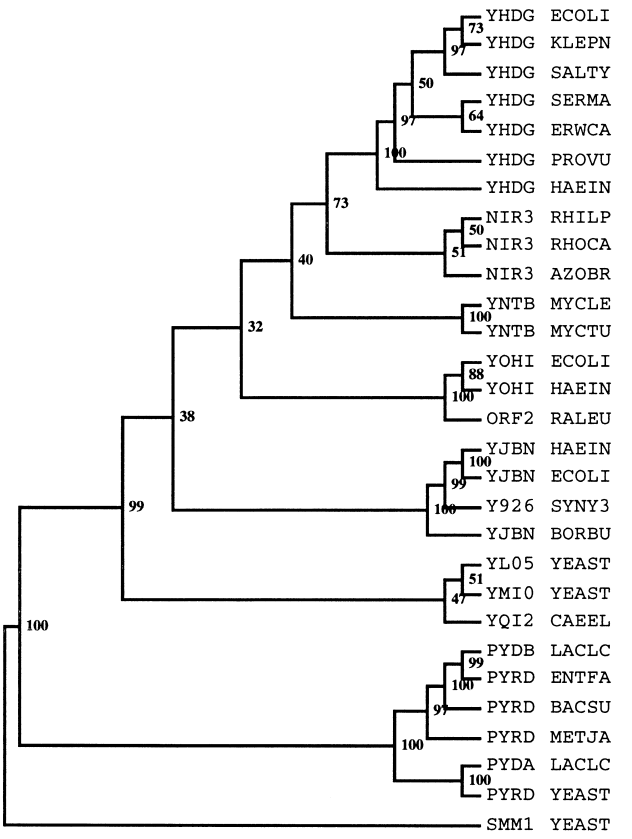

Fig. 4. Consensus parsimony phylogenetic tree of the YhdG and Nifr3 homologue proteins. The multiple sequence alignment of Fig. 3 was utilized for the phylogenetic reconstruction as in Fig. 2. Note that YhdG and NifR3 clustered together.

3.4. Model for the transfer of yhdG-fis

Fig. 5 shows the genetic organization of the *yhdG-fis* and *nifR3-ntrBC* operons in γ and α proteobacterial species respectively. Due to the sequence similarity of YhdG and Fis with NifR3 and the COOH-terminal domain on NtrC, respectively, it is likely that the former genes originated from the latter and were acquired by the γ proteobacterial species by horizontal gene transfer, as discussed above. There are not only structural but also some functional similarities of NtrC and Fis. Both proteins bind to enhancer sites that function in a position-independent manner and form high-order nucleoprotein complexes, although to promote quite different processes and

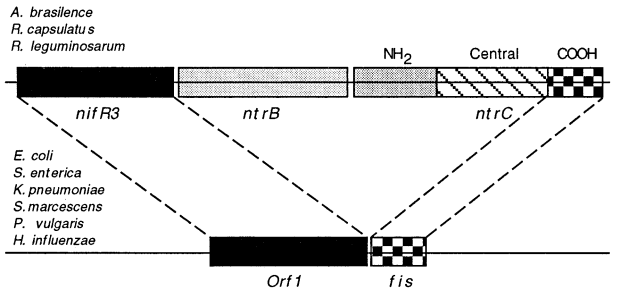

Fig. 5. Genetic organization of the *yhdG-fis* and *nifR3-ntrBC* operons of γ and α proteobacterial species respectively. Homologous sequences (*yhdG* with *nifR3*, and *fis* with the COOH-terminal domain of *ntrC*) are joined by dashed lines.

interacting with unrelated proteins [31,32]. Since Fis is involved in phage insertion and excision and in transposition it is tempting to speculate that the COOH-terminal domain of NtrC was first recruited by a phage, a transposon or an insertion sequence, perhaps by allowing the formation of a more efficient nucleoprotein complex required either for insertional recombination and for DNA replication. Subsequently from one of these vectors *yhdG* and *fis* were transferred to an ancestral γ proteobacterial host.

A recent analysis indicates that gene order and even operon structure is rarely conserved in species as distant as α and γ proteobacteria, unless their products physically interact with each other [33]. The linkage of the *yhdG* and *fis* genes suggests that their products may interact with each other to carry out an as yet unknown function.

From the phylogenetic analysis shown above it is clear that Fis is a relatively recent gene of the γ proteobacterial species. It is remarkable that it became involved in so many bacterial processes in a rather short evolutionary time. The relaxed specificity for DNA recognition [16] could have allowed Fis to interact with many different sites to allow transcriptional activation and to enhance recombination.

3.5. Conclusions

From our phylogenetic analysis of Fis and YhdG we conclude that these genes were acquired by a lineage ancestral to the present day γ proteobacterial species from the *nifR3-ntrBC* operon of an ancestral α proteobacterial lineage by horizontal gene transfer. Fis has evolved to be a highly versatile protein involved in many physiological processes. Thus, this is a clear case of a gene with a new function whose origin can be traced back to a single domain of another protein.

There are multiple reports of horizontally inherited genes in bacteria [1,6]. However, only in a few cases related to the transfer of drug resistance genes or pathogenicity islands between closely related group of strains, the original donor organism or gene have been identified [1,7]. Having complete genome sequence information and using sensitive sequence analysis and phylogenetic tools, the origin of gene function in bacteria can be traced to enhance the understanding of evolution and speciation.

Acknowledgements: We are grateful to Martijn Huynen for his comments on the manuscript. E.M.'s work was funded in part by CONACyT and DGAPA/UNAM. E.M. is an Alexander von Humboldt Fellow.

References

[1] Syvanen, M. (1994) Annu. Rev. Genet. 28, 237–261.
[2] Lawrence, J.G. and Roth, J.R. (1996) Genetics 143, 1843–1860.
[3] Médigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) J. Mol. Biol. 222, 851–856.
[4] Lawrence, J.G. and Ochman, H. (1997) J. Mol. Evol. 44, 383–397.
[5] Ochman, H. and Lawrence, J.G. (1996) in: F.C. Neidhardt, R. Curtis III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter and H.E. Umbarger (Eds.), *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, 2nd Edn., American Society for Microbiology, Washington, DC, pp. 2627–2637.
[6] Smith, M.W., Feng, D.F. and Doolittle, R.F. (1992) Trends Biochem. Sci. 17, 489–493.
[7] Bäumler, A.J. (1997) Trends Microbiol. 5, 318–332.
[8] Finkel, S.E. and Johnson, R.C. (1992) Mol. Microbiol. 6, 3257–3265.
[9] Johnson, R.C., Bruist, M.F. and Simon, M.I. (1986) Cell 46, 531–539.
[10] Koch, C. and Kahmann, R. (1986) J. Biol. Chem. 261, 15673–15678.
[11] Ball, C.A. and Johnson, R.C. (1991) J. Bacteriol. 173, 4027–4031.
[12] Weinreich, M.D. and Reznikoff, W.S. (1992) J. Bacteriol. 174, 4530–4537.
[13] Gille, H., Egan, J.B., Roth, A. and Messer, W. (1991) Nucleic Acids Res. 19, 4167–4172.
[14] Filutowicz, M., Ross, W., Wild, J. and Gourse, R.L. (1992) J. Bacteriol. 174, 398–407.
[15] Kostrewa, D., Granzin, J., Koch, C., Choe, H.W., Raghunathan, S., Wolf, W., Labahn, J., Kahmann, R. and Saenger, W. (1991) Nature 349, 178–180.
[16] Hengen, P.N., Bartram, S.L., Stewart, L.E. and Schneider, T.D. (1997) Nucleic Acids Res. 25, 4994–5002.
[17] Gonzalez-Gil, G., Bringmann, P. and Kahmann, R. (1996) Mol. Microbiol. 22, 21–29.
[18] Osuna, R., Lienau, D., Hughes, K.T. and Johnson, R.C. (1995) J. Bacteriol. 177, 2021–2032.
[19] Johnson, R.C., Ball, C.A., Pfeffer, D. and Simon, M.I. (1988) Proc. Natl. Acad. Sci. USA 85, 3484–3488.
[20] North, A.K., Klose, K.E., Stedman, K.M. and Kustu, S. (1993) J. Bacteriol. 175, 4267–4273.
[21] Morett, E. and Segovia, L. (1993) J. Bacteriol. 175, 6067–6074.
[22] Ninnemann, O., Koch, C. and Kahmann, R. (1992) EMBO J. 11, 1075–1083.
[23] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A. and Merrick, J.M. et al. (1995) Science 269, 496–512.
[24] Beach, M.B. and Osuna, R. (1997) unpublished. Accesion numbers: EMBL:AF040378, EMBL:AF040379, EMBL:AF040380, EMBL:AF040381.
[25] Doerks, T., Bairoch, A. and Bork, P. (1998) Trends Genet. 14, 248–250.
[26] Foster-Hartnett, D., Cullen, P.J., Gabbert, K.K. and Kranz, R.G. (1993) Mol. Microbiol. 8, 903–914.
[27] Patriarca, E.J., Riccio, A., Tate, R., Colonna-Romano, S., Iaccarino, M. and Defez, R. (1993) Mol. Microbiol. 9, 569–577.
[28] Machado, H.B., Yates, M.G., Funayama, S., Rigo, L.U., Steffens, M.B., Souza, E.M. and Pedrosa, F.O. (1995) Can. J. Microbiol. 41, 674–684.
[29] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zheng, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[30] Ochman, H. and Wilson, A.C. (1987) J. Mol. Evol. 26, 74–86.
[31] Reitzer, L.J. and Magasanik, B. (1986) Cell 45, 785–792.
[32] Wyman, C., Rombel, I., North, A.K., Bustamante, C. and Kustu, S. (1997) Science 275, 1658–1661.
[33] Huynen, M. and Bork, P. (1998) Proc. Natl. Acad. Sci. USA 95, 5849–5856.
[34] Thompson, J.D., Higgins, D.G. and Gibson, T. (1994) Nucleic Acids Res. 22, 4673–4680.