Lateral Transfer of an *EF-1*α Gene: Origin and Evolution of the Large Subunit of ATP Sulfurylase in Eubacteria

Yuji Inagaki,[1,3] W. Ford Doolittle,[1]
Sandra L. Baldauf,[2] and Andrew J. Roger[1]
[1]Program in Evolutionary Biology
Canadian Institute for Advanced Research
Department of Biochemistry and Molecular Biology
Dalhousie University
Halifax, Nova Scotia B3H 4H7
Canada
[2]Department of Biology
University of York
Heslington, York YO10 5DD
United Kingdom

Summary

It is generally accepted that new genes arise via duplication and functional divergence of existing genes, in accordance with Ohno's model [1], now called "Mutation During Redundancy," or MDR [2]. In this model, one of the two gene copies is free to acquire novel (although likely related) activities through mutation, since only one copy is required for its original function. However, duplication within a genome is not the only process that might give rise to this situation: acquisition of a functionally redundant gene by lateral gene transfer (LGT) could also initiate the MDR process. Here we describe a probable instance, involving LGT of an archaeal or eukaryotic *elongation factor 1*α (*EF-1*α) gene. The large subunit of ATP sulfurylase (CysN or the N-terminal portion of NodQ), found mainly in proteobacteria, is clearly related to translation elongation factors [3, 4]. However, our analyses show that *cysN* arose from an *EF-1*α gene initially acquired by LGT, not from a within-genome duplication of the resident *EF-Tu* gene. To our knowledge, this is the first unequivocal case of LGT followed by functional modification to be described; this mechanism could be a potentially important force in establishing genes with novel functions in genomes.

Results and Discussion

In *Escherichia coli*, *cysD* and *cysN* encode the two subunits of an ATP sulfurylase that produces adenosine-5′-phosphosulfate (APS) from ATP and sulfate, coupled with GTP hydrolysis [5]. APS is then phosphorylated by an APS kinase, encoded by *cysC*, to produce 3′-phosphoadenosine-5′-phosphosulfate (PAPS), which is then used in amino acid (aa) biosynthesis [5]. On the other hand, the Rhizobiaceae group (α-proteobacteria) appear to carry out the same chemistry for the sufation of a nodulation factor. In *Rhizobium meliloti*, a heterodimeric complex comprised of NodP and NodQ appears to possess ATP sulfurylase and APS kinase activities [6]. Indeed, NodP shows strong aa sequence similarity to CysD, while

[3]Correspondence: yinagai@is.dal.ca

nodQ appears to encode both CysN- and CysC-related sequences in a single open reading frame (ORF) (the N and C termini of NodQ correspond to CysN and CysC, respectively) [3].

To date, *cysN* and *nodQ* genes have been reported only from proteobacteria and a few gram-positive bacteria (*Mycobacterium tuberculosis*, *Streptomyces coelicolor*, and *Clostridium acetobutylicum*). Other eubacteria, Archaea, and eukaryotes use a different ATP sulfurylase, which shows no aa sequence similarity to CysN or NodQ. Instead, aa sequences of CysN and the N-terminal portion of NodQ (henceforth referred to as CysN/NodQ*) do show similarity to GTPases involved in translation—in particular, EF-Tu and EF-1α [3, 4]. However, when CysN or NodQ sequences are used as a query, the top BLASTP [7] matches against the GenBank nonredundant database are archaeal EF-1α sequences. An alignment including CysN/NodQ*, EF-1α, and EF-Tu contains large regions of ambiguously aligned and/or gap-containing positions, and only 141 positions can be aligned among all three families, corresponding to ~30%–35% of the lengths of the protein sequences (length range, ~400–470 residues).

Over those positions, CysN/NodQ* is more similar to EF-1α than to EF-Tu; CysN/NodQ* has an average identity of 44% and 42% to archaeal and eukaryotic EF-1α, respectively, whereas it is 36% identical to EF-Tu (see the Supplementary Material available with this article online). Furthermore, the entire aa sequences of archaeal or eukaryotic EF-1α can be aligned to CysN/NodQ* without any of the large gaps that are required for an alignment of EF-Tu with EF-1α or with CysN/NodQ* sequences. A total of 344 unambiguously aligned positions can be extracted from the EF-1α+CysN/NodQ* alignment (203 more than the EF-Tu+EF-1α+CysN/NodQ* alignment), corresponding to ~75%–85% of the lengths of the protein sequences.

Of particular interest is a large insertion/deletion (indel) in the N-terminal region of EF-1α shown in Figure 1A. Archaeal and eukaryotic EF-1α share this insertion, which is absent from EF-Tu or their ancient paralog EF-G/2. In the tertiary structures of EF-1α from yeast and *Sulfolobus*, the regions corresponding to this insertion form α helices (Figure 1A). Interestingly, the CysN/NodQ* family also has an insertion in this position that is similar in length to the EF-1α sequences (Figure 1A), although it has little aa sequence similarity to EF-1α sequences over this region. Nevertheless, the central region in the CysN/NodQ* insertion is predicted to form a putative α helix by two different protein secondary structure prediction programs (Figure 1A) in the region corresponding to the EF-1α α2 helix. This suggests that these insertions are homologous. If so, this insertion, coupled with the pairwise sequence identities and overall alignment lengths would suggest that the CysN/NodQ* family had directly evolved from either an archaeal or eukaryotic EF-1α, not from eubacterial EF-Tu.

Phylogenetic analyses also support a relatively strong evolutionary affinity between CysN/NodQ* and EF-1α to

A   **B**


```
                              19                                                              71   112          134
                              |                                                               |    |            |
                              +++++                                             +++++++++++++      +++++++    +++
         Aeropyrum pernix     GKTTTS DSLLAAAGIISERIAGE-- ----------------------------------- ALVLVDYLNVEKQRGI  AIVVVDAVE-------GVMTQTE
EF-G/2   Halobacterium halobium GKTTLT DNLLAGAGMISEDTAGQ-- --------------------------------- QLAMDTEEDEQERGI  ALVVVDAVE-------GAMPQTE
         Dictyostelium discoideum GKTTLS DSLIQRAGIIADKVSGD-- ------------------------------- MRYMSCRADEQERGI  ALVVIDCVE-------GVCVQTE
         Thermus thermophilus  GKTTTT ERILYYTGRIHKIGEVHEG ----------------------------------- AATMDFMEQERERGI  AIVVFDSSQ-------GVEPQSE
         α helix (1FNM)        [  α1  ]

         Mycoplasma genitalium GKTTLT -----AAICTVLAKEGKSAATR-- ------------------------------- YDEIDKAPEEKARGI  AILVVSATD-------SVMPQTR
         Mycbacterium leprae   GKTTLT -----AAITKVLHDKFPNLNESRA ------------------------------- FDQIDNAPEERQRGI  AILVVAATD-------GPMPQTR
EF-Tu    Haemophilus influenzae GKTTLT -----AAITTVLAKHYGGAARA-- ------------------------------ FDQIDNAPEEKARGI  AILVVAATD-------GPMPQTR
         Neisseria gonorrhoeae GKTTLT -----AALTTILAKKFGGAAKA-- ------------------------------ YDQIDNAPEEKARGI  AILVCSAAD-------GPMPQTR
         Rickettsia prowazekii GKTSLT -----AAITIILAKTGGAKATA-- ------------------------------ YDQIDAAPEEKARGI  AILVVSAAD-------GPMPQTR
         Thermus aquaticus     GKTTLT -----AALTYVAAAENPNVEVKD- ------------------------------ YGDIDKAPEERARGI  AILVVSAAD-------GPMPQTR
         α helice (1B23)       [===== α1 =====]                                              [  α2  ]

         Haloarcula marismortui GKSTLV ----- GRLLYETGSVPEHVIEQHKEEAEEKGKG--GPEF AYVMDNLAEERERGV  AVLVVAADD-------GVQPQTQ
         Pyrococcus horikoshii GKSTTI ----- GRLLYDTGNIPETII-KKFEEMGEKGK---SFKP AWVMDRLKEERERGI  AVLVVAATD-------GVQPQTR
archaeal Archaeoglobus fulgidis GKSTLI ----- GRLLYETGEIPEHIIEKMRKEAQEKGKA--TFEF AWVMDRLKEERERGI  AVLVMDVVE-------KVQPQTR
EF-1α    Desulfurococcus mobilis GKSTMT ----- GHLIYRLGYFDEKTVKMIEEESKKMGKE--SFKP AWLLDRMKEERERGV  AILVVSARK-GEFEAGMSAEGQTR
         Aeropyrum pernix      GKSTLV ----- GHLLYRLGYIEEKKLKELEEQAKSRGKE--SFKF AWILDKMKEERERGI  AILVVSARK-GEFEAGMSTEGQTR
         Sulfolobus solfataricus GKSTLV ----- GRLLMDRGFIDEKTVKEAEEAAKKLGKE--SEKF AFLLDRLKEERERGV  AILVVSAKK-GEYEAGMSVEGQTR
         α helice (1JNY)       [===== α1 =====]      [ α2 ]                      [  α3  ]

         Euglena gracilis      GKSTTT ----- GHLIYKCGGIDKRTIEKFEKEASEMGKG--SFKY AWVMDKLKAERERCI  AVLVIDSTT-GGFEAGISKDGQTR
         Dictyostelium discoideum GKSTTT ----- GHLIYKCGGIDKRVIEKYEKEASEMGKQ--SFKY AWVMDKLKAERERGI  AVLVIASPT-GEFEAGIAKNGQTR
eukaryotic Porphyra purpurea   GKSTTT ----- GHLIYKCGGIDKRVIEKFEKEAAEMGKG--SFKY AWVLDKLKAERERGI  AILVIASPP-GEFEAGISQNGQTR
EF-1α    Arabidopsis thaliana  GKSTTT ----- GHLIYKLGGIDKRVIERFEKEAAEMNKR--SFKY AWVLDKLKAERERGI  AVLIIDSTT-GGFEAGISKDGQTR
         Saccharomyces cerevisiae GKSTTT ----- GHLIYKCGGIDKRTIEKFEKEAAELGKG--SFKY AWVLDKLKAERERGI  AILIIAAGV-GEFEAGISKDGQTR
         α helice (1IJF)       [===== α1 =====]      [ α2 ]                      [  α3  ]

         Neisseria meningitidis GKSTLI ----- GRLLYDSKALLSDQIKTLESGKSKGLK--ETIDF SILTDGLEAEREQGI  AVVLVDASQLDFGAQPLQLLPQTK
         Allochromatium vinosum GKSTLI ----- GRLLYDTKAILADTLHAIAATSQRRGL--SELDL SLLTDGLQAEREGGI  AIILVDARR-------GVQTQTR
         Vibrio cholerae       GKSTLI ----- GRLLHDSKQIYEDQLAAVHNDSQRVGTTGSRPDL ALLVDGLQARREQGI  AVILIDARK-------GVLDQTR
         Streptomyces coelicolor GKSTLV ----- GRLLHDSKSVLTDQLEAVERASASRGQ--DVPDL ALLTDGLRAEREQGI  TVILVDARN-------GVVEQTR
         Klebsiella aerogenes  GKSTLI ----- GRLLHDTRQIYEDQLSSLHNDSKRHGTQGEKIDL ALLVDGLQAEREQGI  AILLIDARK-------GVLDQTR
         Escherichia coli      GKSTLI ----- GRLLHDTKQIYEDQLSSLHNDSKRHGTQGEKLDL ALLVDGLQAEREQGI  AILLIDARK-------GVLDQTR
         Buchnera sp.          GKSTLI ----- GRLLHDTKQIYDDQLFFLKSDSKRHGTQGNEIDL ALVVDGLQSEREQGI  SILLVDARK-------GLSEQTY
         Agrobacterium tumefaciens GKSTLI ----- GRLLWDTKAVKEDQAATLHRDSGKQNDLG-LPDF ALLLDGLQAEREQGI  AVLLVDART-------GILEQTR
         Sinorhizobium meliloti GKSTLI ----- GRLLWDTKAVKEDQAASLQRDSSGKQNDLGLPDF ALLLDGLQAEREQGI  AVLLVDARV-------GLLEQTR
CysN     Xylella fastidiosa    GKSTLI ----- GHLLYDSQCLAEDQLADLMVDSQRYGTQGEHIDY ALLLDGLAAEREQGI  AVVLVDARK-------GLLTQTR
+NodQ    Rhizobium tropici     GKSTLI ----- GRLLCDAQLVFEDQLADLRHGGIRGGNG-EEIDF SLVLDGLEAEREQGI  AVILVDSRQ-------GILQQTR
         Rhizobium sp. BR816   GKSTLI ----- GRLLVDAKLVFEDQLTNLGRVGSSGAVNGGEIDL ALLLDGLEAEREQGI  AVILIDSRQ-------GILQQTR
         Rhizobium sp. N33     GKSTLI ----- GRLLSDTKQIFEDQLAALERDSRKHGTTGDDIDF ALLVDGLEAEREQGI  AIVLIDARQ-------GVLRQTR
         Pseudomonas aeruginosa GKSTLI ----- GRLLHDSKMIYDHLEAITRDSKKVGTTGDDVDL ALLVDGLEAEREQGI  AIILIDARY-------GVQTQTR
         Caulobacter crescentus GKSTLI ----- GRLLYDSKMIFEDQMAALEVDSRKVGTQGGAIDF ALLVDGLAAEREQGI  AVILVDARH-------GVLTQTR
         Mycobacterium tuberculosis GKSTLI ----- GRLLYDSKAVMEDQWASVEQTKSDRGHD--YTDL ALVTDGLRAEREQGI  VIVLVDARH-------GLLEQSR
         Rhizobium meliloti pSym GKSTLI ----- GRLLYDAKLVFEDQLANLGRVGSPGAANGKEIDL ALLLDGLEAEREQGI  AIILIDSRQ-------GILQQTR
         Rhizobium loti        GKSTLI ----- GRLLSDTKQIFEDQLAALERDSRKHGTTGDDIDF ALLVDGLEAEREQGI  AIVLIDARQ-------GVLRQTR
         Azospirillum brasilense GKSTLI ----- GRLLHDAGLISDDQLEQARRDSRGRAEEDGGIDF SLLVDGLEAEREQGI  AVLLVDARH-------GLLTQTR
         Clostridium acetobutylicum GKSTLI ----- GRLLYDTNSLPDGAIEKVKKISAEEGK---KPEY AFLLDAPEEEQRQGI  AILVVDAKE-------GIQEQSK
         Potential helix (SOPA) HHH                     HHHhhHHHHhhhHHHHHHHHHHh       hH hhhHHHHHHHHhh
         Potential helix (nnPredict)                    h HH  hhHHHHHHH              Hh hHHh  HHhH
```
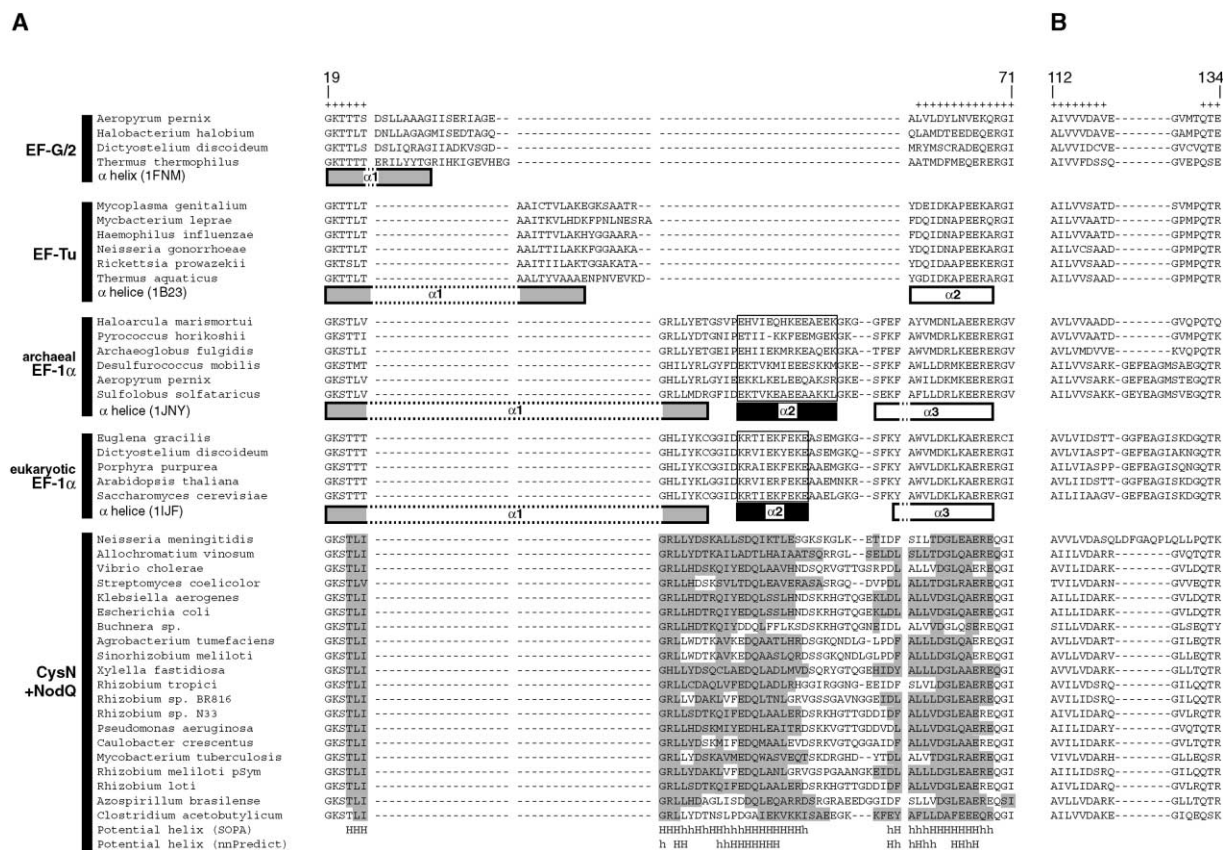
Figure 1. Insertion/Deletions in EF-Tu, EF-1α, CysN/NodQ*, and EF-G/2

(A) Amino acid alignment that corresponds to position 19 to 71 in *Saccharomyces cerevisiae* EF-1α (GenBank accession number CAA25356). The observed α helices in *Thermus thermophilus* EF-G (PDB accession number 1FNM), *T. aquaticus* EF-Tu (1B23), *Sulfolobus solfataricus* EF-1α (1JNY), and *Saccharomyces* EF-1α (1IJF) are indicated by boxes (structurally homologous helices are shown in the same colors). The region corresponding to the α2 helices of *Saccharomyces* and *Sulfolobus* EF-1α are boxed. The CysN/NodQ* insertion and its adjacent region (position 19–71 in *Saccharomyces* EF-1α) predicted to form a putative α helices are shaded. Putative secondary structures in the CysN/NodQ* insertion were predicted by SOPM with the default settings (npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopm.html) [21]. The positions predicted to form helices in at least 50% or at least 80% of CysN/NodQ* sequences are highlighted by "h" or "H," respectively (first line below the alignment). We also predicted helices by nnPredict with the default settings (www.cmpharm.ucsf.edu/~nomi/nnpredict.html) [22], and only the consensus results are shown on the second line below the alignment (criteria as described above).

(B) Amino acid alignment that corresponds to position 112–134 in *Saccharomyces* EF-1α. + indicates the positions used for the phylogenetic analyses.

the exclusion of EF-Tu (Figure 2A). We analyzed a large data set including EF-Tu, EF-1α, CysN/NodQ*, and EF-G/2 (64 sequences with 90 sites), using distance and maximum likelihood (ML) methods incorporating a gamma (Γ) distribution model for among-site rate variation. The compensation for among-site rate variation in these analyses is important, since this data set included highly diverged "long branched sequences" (CysN/NodQ* and EF-G/2 in particular). Such long branches can artifactually "attract" one another in phylogenetic analyses, especially when among-site rate variation is not taken into account (e.g., see [8]). The CysN/NodQ* and EF-1α sequences grouped together to the exclusion of the EF-Tu and EF-G/2 sequences with bootstrap percent support (BP) = 85% in the distance tree (Figure 2A). Due to the computational intensity of full ML analysis, we generated a smaller data set (23 sequences) using the same alignment and analyzed it by the ML method (Figure 2A). A similar result was obtained from the ML analyses, albeit with somewhat poorer support: BP = 61% (Figure 2A).

To recover more phylogenetic resolution, we analyzed data sets including only archaeal/eukaryotic EF-1α and CysN/NodQ* (42 sequences with 344 sites). Although the exclusion of divergent EF-Tu and EF-G/2 sequences makes an additional 254 positions available for the analyses, the phylogeny fails to clarify whether Archaea, eukaryotes, or a lineage that split off earlier from their common ancestor were the donors of the EF-1α that gave rise to CysN/NodQ* (see Supplementary Material). However, another indel in the alignment provides an important clue for the origin of the CysN/NodQ* family. It is well known that crenarchaeotes and eukaryotes share a seven-residue insertion in EF-1α that is absent from euryarchaeal homologs (Figure 1B) [9, 10]. With the exception of *Neisseria*, the CysN sequences, like the euryarchaeotes, lack an insertion at this position. Furthermore, the insertion sequence in the *Neisseria* CysN is not similar to those in crenarchaeal/eukaryotic EF-1α (Figure 1B) and therefore is unlikely to be homologous. The simplest interpretation of these data is that
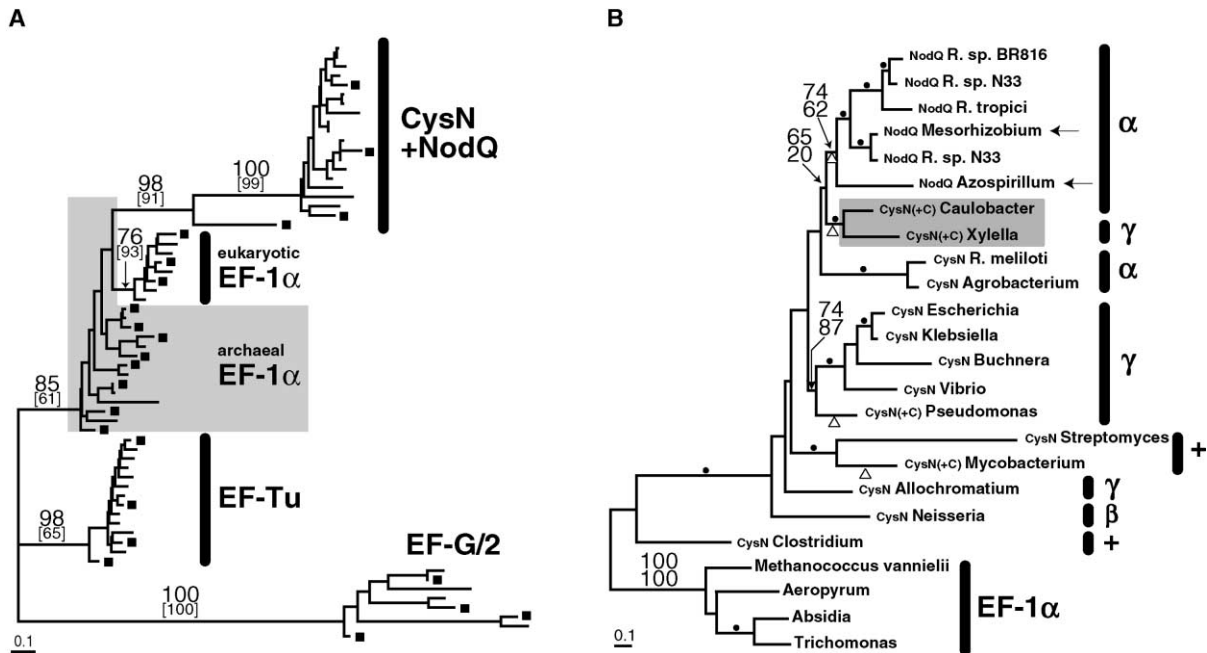
A

B

Figure 2. Phylogenetic Analyses of CysN/NodQ*

(A) Phylogenetic position of the CysN/NodQ* family in the EF superfamily. We manually added CysN/NodQ* sequences to a previously published EF-Tu+EF-1α+EF-G/2 alignment [10]. Fourteen EF-Tu and 22 EF-1α sequences are comprehensively sampled from eubacteria, Archaea, and eukaryotes. Two EF-G/2 sequences are chosen from the eubacteria, Archaea, and eukaryotes. All 20 CysN/NodQ* sequences available in the GenBank databases as of October 2001 were considered in this study. After exclusion of ambiguously aligned and gapped sites, 90 sites shared by the five protein families (total of 64 sequences with sites used corresponding to positions 9-24, 58-74, 87-118, and 132-157 in *Absidia glauca* EF-1α [GenBank accession number P2829], accounting for ~20% of length of the protein sequence) remained for the phylogenetic analyses. A distance matrix was calculated with the JTT aa substitution model incorporating among-site rate variation (JTT + Γ model) (discrete Γ distribution, eight categories) using TREE-PUZZLE v. 4.0.2 [23]. A tree was reconstructed from the distance matrix, using FITCH with global rearrangements and ten jumbles implemented in PHYLIP v. 3.5 [24]. The bootstrap analysis with the FITCH method employing the JTT + Γ model (100 replicates) was carried out using TREE-PUZZLE, PUZZLEBOOT v. 1.02 (A.J. Roger and M.E. Holder; members.tripod.de/korbi/puzzle/), SEQBOOT, FITCH, and CONSENSE in PHYLIP. Bootstrap percentage support values (BP%) are only indicated for major nodes. The shape parameter α = 0.87 estimated from the data by TREE-PUZZLE was used throughout this series of analyses. The identical tree with full sequence names and accession numbers are described in the Supplementary Material. A smaller data set that included four representative CysN/NodQ*, 11 EF-1α, four EF-Tu, and four EF-G/2 sequences (total 23 sequences marked by filled boxes) was used for the bootstrap analysis with the ML method employing the PAM + Γ model (discrete Γ distribution, four categories) using PROML in the PHYLIP v. 3.6 [24] (BP% are given in brackets). Note that the relatively large discrepancies between bootstrap values inferred with ML versus distance methods from different taxonomic samples of the same alignment are not unexpected [25].

(B) Phylogenetic relationships among CysN/NodQ* sequences. All 20 CysN/NodQ* sequences available as of October 2001 were analyzed with two archaeal and two eukaryotic EF-1α sequences as outgroups. The alignment includes 24 sequences with 344 unambiguously aligned sites (no gapped positions used; corresponding to positions 21-39, 41-42, 44-51, 53-118, 132-181, 189-201, 205-211, 224-241, 243-256, 259-260, 263-282, 286-316, 319-334, 339-349, 351-378, and 389-425 in *Absidia* EF-1α, accounting for ~75% of length of the protein). The tree was reconstructed by using PROML as described above, except eight discrete Γ distribution categories were used. The shape parameter α = 1.05 was estimated from the data by TREE-PUZZLE. Bootstrap analyses (100 replicates) were carried out with PROML and FITCH methods as described above. The branches that received BP > 90% in bootstrap analyses are highlighted by dots. Triangles indicate putative fusion events of *cysN* and *cysC* genes. The *cysN* genes that fused with *cysC* are indicated as "CysN(+C)." The *Xylella* + *Caulobacter* clade is shaded, since this clade is likely the result of a secondary LGT. Plasmid-encoded *nodQ* genes are marked by arrows. Eubacterial classifications are given in brackets; α, β, γ, and + indicate α-, β-, and γ-proteobacteria and gram-positive bacteria, respectively.

CysN/NodQ* evolved by LGT from an *EF-1α* gene in the euryarchaeal or a pre-eukaryotic-archaeal lineage and that the lineage leading to *Neisseria* CysN acquired an insertion independently from the crenarchaeote/eukaryote EF-1α lineages. Likewise, it is unlikely that CysN/NodQ* is directly related to the eukaryote-specific EF-1α paralogs, eukaryotic release factor 3 and HBS1, as these two protein families also include the crenarchaeote-eukaryote insertion [11].

A detailed CysN/NodQ* phylogeny using a data set with 20 CysN/NodQ* and four EF-1α sequences as outgroups (344 unambiguously aligned sites) reveals a putative LGT of *cysN* within eubacteria (Figure 2B). Five

out of seven γ-proteobacterial CysN sequences form a robust clade (Figure 2B). However, the CysN sequences of the two γ-proteobacteria, *Xylella* and *Allochromatium*, do not join this clade. While the *Allochromatium* CysN shows no evolutionary affinity to other sequences BP > 50%, the *Xylella* sequence robustly groups with that from an α-proteobacterium, *Caulobacter* (Figure 2B), and so these eubacteria likely exchanged a fused *cysNC* gene. However, the direction of this LGT is unclear due to the low phylogenetic resolution. Alternatively, an as yet unknown lineage might have donated the gene to both *Xylella* and *Caulobacter* genomes independently.
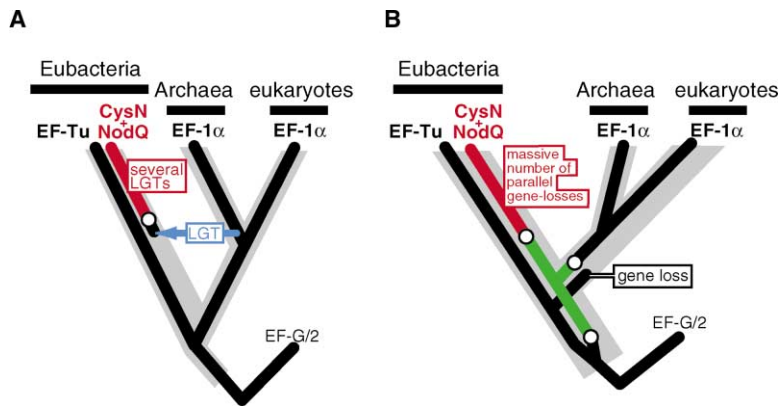
Figure 3. Alternative Hypotheses for the Origin of CysN/NodQ*

The gene phylogenies for the preferred LGT scenario (A) and the alternative deep paralogs scenario (B) are drawn schematically. Open circles indicate functional conversions along the CysN/NodQ* lineage. Line colors represent the biochemical functions: black, EF; green, an unknown (non-EF) function; and red, CysN/NodQ*. The organismal relationships among Eubacteria, Archaea, and eukaryotes are shown by shading.

NodQ is a fusion protein of CysN and CysC that was modified for the molecular machinery of nodule formation. All NodQ* sequences form a monophyletic clade, and that emerges from within the α-proteobacterial group of CysN sequences in the ML tree (Figure 2B). Therefore, we suggest that a fusion event of *cysN* and *cysC* followed by functional modification gave rise to the ancestral *nodQ* gene. Our analyses revealed three additional *cysN-cysC* fusion events. Although the *Caulobacter + Xylella* CysNC clade was adjacent to the NodQ* clade in the ML tree, the branch that connects these groups was not strongly supported by bootstrap analyses (Figure 2B). The *cysN-cysC* fusion in the *Caulobacter + Xylella* lineage therefore could be independent from the origin of NodQ. The fusion events along the branches leading to *Pseudomonas* and *Mycobacterium* must be also independent, since the two sequences branch with standard CysN sequences from γ-proteobacteria and *Streptomyces*, respectively (Figure 2B). These multiple independent fusions may betray a selective advantage that accrues to fused *cysN-cysC* genes. Once CysN and CysC are fused, the two "proteins" that catalyze sequential reactions in sulfur metabolism would be highly concentrated in a limited space, as will their products and substrates, and thus their enzymatic activities may be enhanced [12, 13]. Such fusion events could also be selectively neutral, resulting from mutations in a spacer region between *cysN* and *cysC* that convert the two protein-coding genes into single ORF.

Although the "LGT" scenario for the CysN/NodQ* evolution is most likely, an alternative explanation—referred to here as the "deep paralogs" scenario—is possible (schematically drawn in Figure 3). Two versions of translation EFs (i.e., EF-Tu and the ancestral molecule of CysN and EF-1α, here referred to as "CysN*-EF-1α") could have been established in the last universal common ancestor(s) (known as the *cenancestor* or *cenancestors*, since there was likely a population of such entities). Subsequently, after the divergence of the eubacteria from the archaeal-eukaryotic lineage, one of these EFs, CysN*-EF-1α, could have been transformed into ATP sulfurylase during eubacterial evolution. In this scenario, the ancestor of Archaea and eukaryotes lost the EF-Tu paralog, retaining only the CysN*-EF-1α paralog as its EF (Figure 3B).

However, the deep paralogs scenario is unlikely for several reasons. The phylogenetic distribution of *cysN* or *nodQ* genes is restricted to proteobacteria and a few gram-positive bacteria. Thus, if the deep paralogs scenario were correct, massive numbers of parallel gene losses of CysN*-EF-1α must be invoked to explain this restricted phylogenetic distribution within eubacteria (Figures 3B). Furthermore, it is unlikely that the cenancestor had redundant EFs, and either EF-Tu or CysN*-EF-1α was the principal EF in the cenancestors (the other would have had a non-EF function). If so, a switch of the EF function between CysN*-EF-1α and EF-Tu would have had to take place either in the eubacterial or archaeal-eukaryotic lineage (Figure 3B). In contrast, the LGT scenario requires no transfer of EF function between paralogs, only a single LGT event to establish the *cysN*-EF-1α gene in a eubacterium followed by functional diversification. Subsequently, the *cysN* gene was inherited and/or spread via secondary LGT events into other eubacterial lineages (Figure 3A). The presence of such genes on plasmids [14, 15] and their "scrambled" phylogenetic distribution among eubacteria is consistent with the "evolutionary mobility" of *cysN* genes.

LGT is now realized as a principal mechanism by which prokaryotic genomes diversify [16]. In well known LGT cases, the function of the gene before and after the gene transfer event remains the same (e.g., see [17, 18]). However, the origin of the CysN/NodQ* family in eubacteria via an interdomain LGT of an EF-1α, if correct, represents a new twist in our understanding of the protein evolution. The traditional MDR model [1, 2] needs to be updated to accommodate functional diversification after LGT. As with the MDR model, it will be important to determine how functionally identical duplicates can escape from frequent silencing mutations until one of the duplicates acquires rare advantageous mutations [19]. In any case, as Lawrence and Roth have argued [20], the prevalence of LGT among prokaryotes and the "quantum" leaps over sequence space it permits (in contrast to point mutation) suggests it could play a more important role in the evolution of gene function than previously recognized.

References

1. Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer-Verlag).

2. Hughes, A.L. (1999). Evolution of new protein function. In Adaptive Evolution of Genes and Genomes. A.L. Hughes, ed. (Oxford, NY: Oxford University Press), pp. 143–179.

3. Cervantes, E., Sharma, S.B., Maillet, F., Vasse, J., Truchet, G., and Rosenberg, C. (1989). The *Rhizobium meliloti* host range *nodQ* gene encodes a protein which shares homology with translation elongation and initiation factors. Mol. Microbiol. *3*, 745–755.

4. Leyh, T.S., Vogt, T.F., and Suo, Y. (1992). The DNA sequence of the sulfate activation locus from *Escherichia coli* K-12. J. Biol. Chem. *267*, 10405–10410.

5. Kredich, K.M. (1996). Biosynthesis of cysteine. In *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, K.M. Kredich, ed. (Washington, D.C.: American Society of Microbiology), pp. 514–527.

6. Schwedock, J.S., and Long, S.R. (1992). *Rhizobium meliloti* genes involved in sulfate activation: the two copies of *nodPQ* and a new locus, *saa*. Genetics *132*, 899–909.

7. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

8. Sullivan, J., and Swofford, D.L. (2001). Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. *50*, 723–729.

9. Rivera, M.C., and Lake, J.A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science *257*, 74–76.

10. Baldauf, S.L., Palmer, J.D., and Doolittle, W.F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA *93*, 7749–7754.

11. Inagaki, Y., and Doolittle, W.F. (2000). Evolution of the eukaryotic translation termination system: origins of release factors. Mol. Biol. Evol. *17*, 882–889.

12. Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86–90.

13. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. Nature *402*, 83–86.

14. Vieille, C., and Elmerich, C. (1990). Characterization of two *Azospirillum brasilense* Sp7 plasmid genes homologous to *Rhizobium meliloti nodPQ*. Mol. Plant Microbe Interact. *3*, 389–400.

15. Galibert, F., Finan, T.M., Long, S.R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P., et al. (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science *293*, 668–672.

16. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304.

17. Boucher, Y., and Doolittle, W.F. (2000). The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. Mol. Microbiol. *37*, 703–716.

18. Nesbo, C.L., L'Haridon, S., Stetter, K.O., and Doolittle, W.F. (2001). Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol. Biol. Evol. *18*, 362–375.

19. Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459–473.

20. Lawrence, J.G., and Roth, J.R. (1999). Genomic flux: genome evolution by gene loss and acquisition. In Organization of Prokaryotic Genome, R.L. Charlebois, ed. (Washington, D.C.: American Society of Microbiology), pp. 263–289.

21. Geourjon, C., and Deleage, G. (1994). SOPM: a self-optimized method for protein secondary structure prediction. Protein Eng. *7*, 157–164.

22. Kneller, D.G., Cohen, F.E., and Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. *214*, 171–182.

23. Strimmer, K., and von Haeseler, A. (1996). Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. *13*, 964–969.

24. Felsenstein, J. (1993). PHYLIP (phylogeny inference package) (Seattle, WA: University of Washington).

25. Roger, A.J., Sandblom, O., Doolittle, W.F., and Philippe, H. (1999). An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes. Mol. Biol. Evol. *16*, 218–233.