Research

Interkingdom gene fusions
Yuri I Wolf, Alexey S Kondrashov and Eugene V Koonin

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Rockville Pike, Bethesda, MD 20894, USA.

Correspondence: Eugene V Koonin. E-mail: koonin@ncbi.nlm.nih.gov

Published: 4 December 2000
Genome Biology 2000, 1(6):research0013.1–0013.13
Received: 5 June 2000
Revised: 11 September 2000
Accepted: 6 November 2000
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/6/research/0013
© Genome Biology.com (Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: Genome comparisons have revealed major lateral gene transfer between the three primary kingdoms of life - Bacteria, Archaea, and Eukarya. Another important evolutionary phenomenon involves the evolutionary mobility of protein domains that form versatile multidomain architectures. We were interested in investigating the possibility of a combination of these phenomena, with an invading gene merging with a pre-existing gene in the recipient genome.

Results: Complete genomes of fifteen bacteria, four archaea and one eukaryote were searched for interkingdom gene fusions (IKFs); that is, genes coding for proteins that apparently consist of domains originating from different primary kingdoms. Phylogenetic analysis supported 37 cases of IKF, each of which includes a ‘native’ domain and a horizontally acquired ‘alien’ domain. IKFs could have evolved via lateral transfer of a gene coding for the alien domain (or a larger protein containing this domain) followed by recombination with a native gene. For several IKFs, this scenario is supported by the presence of a gene coding for a second, stand-alone version of the alien domain in the recipient genome. Among the genomes investigated, the greatest number of IKFs has been detected in Mycobacterium tuberculosis, where they are almost always accompanied by a stand-alone alien domain. For most of the IKF cases detected in other genomes, the stand-alone counterpart is missing.

Conclusions: The results of comparative genome analysis show that IKF formation is a real, but relatively rare, evolutionary phenomenon. We hypothesize that IKFs are formed primarily via the proposed two-stage mechanism, but other than in the Actinomycetes, in which IKF generation seems to be an active, ongoing process, most of the stand-alone intermediates have been eliminated, perhaps because of functional redundancy.

Background
Comparative genome analysis has revealed major lateral gene transfer between the three primary kingdoms of life, Bacteria, Archaea, and Eukarya [1-4]. The best recognized form of lateral gene flux is the transfer of numerous genes from mitochondria and chloroplasts to eukaryotic nuclear genomes [5]. Far beyond that, however, the role of lateral gene exchange, along with lineage-specific gene loss, as one of the principal factors of evolution, at least among prokaryotes, is obvious from the fact that the great majority of conserved families of orthologous genes show a ‘patchy’ phyletic distribution [6,7]. In many cases, such families are shared by phylogenetically distant species (for example, bacteria and archaea), while they are missing in some of the more closely related species (for example, bacteria from the same lineage). Correlations have been noticed between the preferred routes of gene transfer and the lifestyles of the organisms involved. Thus, massive gene exchange seems to have occurred between
archaeal and bacterial hyperthermophiles [8,9], whereas certain parasitic bacteria, for example, chlamydia and spirochetes, appear to have acquired significantly more eukaryotic genes than free-living bacteria [10-12].

Another evolutionary trend that is predominant in eukaryotes, but is important also in bacteria and archaea, involves the evolutionary mobility of protein domains that combine to form variable multidomain architectures [13-18]. Domain fusion is one of the foundations of most forms of regulation and signal transduction in the cell. Examples include prokaryotic transcriptional regulators, most of which consist of the DNA-binding helix-turn-helix domain fused to a variety of small-molecule-binding domains [19], the two-component signal transduction system that is based on fusions of histidine kinases with sensor domains and of receiver domains with DNA-binding domains [20], and the sugar phosphotransferase (PTS) systems that include complex fusions of several enzymes [21]. In the evolution of eukaryotes, domain fusion takes the form of domain accretion, whereby proteins from complex organisms (such as animals) that are involved in various forms of regulation and signal transduction tend to accrue multiple domains that facilitate the formation of complex networks of interactions [22].

We were interested in exploring the possibility of a meeting between these two major evolutionary phenomena - lateral gene exchange and gene fusion - which would result in the formation of multidomain proteins in which different domains display distinct evolutionary provenance. In particular, we sought to identify fusions between domains originating from different primary kingdoms - Bacteria, Archaea and Eukarya - which we term interkingdom gene (domain) fusions (IKFs), and obtain clues to the pathways of IKF origin through comparative genome analysis. We show that, although IKF in general is a rare phenomenon, one bacterial lineage, the Actinomycetes, displays a significantly increased frequency of such events; we also propose a probable mechanism for IKF formation.

**Results and discussion**

To identify IKFs, all protein sequences encoded in the analyzed genomes were compared to the non-redundant protein database, and those proteins in which distinct parts showed the greatest similarity to homologs from different primary kingdoms were identified (see the Materials and methods section). In most cases, the reported alignments were highly statistically significant, leaving no doubt that true homologs were detected (Table 1). On the few occasions when the database search statistics in themselves were not fully convincing (for example, the OB-fold nucleic acid-binding domain in the *Bacillus subtilis* protein YhCN and the methyltransferase domain in the YabN protein, also from *B. subtilis*), the homologous relationship was validated by detection of the salient sequence motifs known to be involved in the corresponding protein functions (data not shown). Such motif analysis was performed for all analyzed domains in order not only to validate homology, but also to distinguish between active and inactivated forms of enzymes. Figure 1 shows multiple alignments of two domains involved in an IKF, illustrating the conservation of the characteristic functional motifs and the specific similarity between each of the domains of the IKF protein (in this case from *Aquifex aeolicus*) and their archaeal and bacterial homologs, respectively.

In several cases, the chimeric origin of a gene was obvious at a qualitative level because no homolog of the ‘alien’ domain with comparable sequence similarity was detected in the recipient superkingdom (Table 1, Figure 2a,b). For the rest of the candidate IKFs, phylogenetic tree analysis was performed to corroborate the origin of the invading domain by horizontal transfer; statistically significant grouping of a candidate IKF domain with homologs from the donor superkingdom provides such evidence (Figure 2c,d). The overall number of confirmed IKFs is relatively small - 37 in 21 compared genomes (about 0.1% of the genes) - compared to the total number of likely interkingdom gene transfers. For completely sequenced bacterial genomes this has been conservatively estimated as 1-2% of the genes, with a greater fraction (2-10%) detected in archaea and hyperthermophilic bacteria ([23]), and K.S. Makarova, L. Aravind and E.V.K., unpublished observations). Examination of the clusters of orthologous groups (COGs) of proteins from complete genomes [6], in which multidomain proteins are split into the constituent domains if the orthologs of the latter are present as stand-alone forms in some of the genomes, shows that IKFs constitute only a small fraction of all fusions of

---

**Figure 1**

Multiple alignments of two domains comprising an interkingdom domain fusion. Alignments of (a) the PHP-hydrolase domain [4] and (b) the pyruvate formate lyase activating enzyme domain of the IKF protein aq_2060 from *A. aeolicus*. The sequences of the aq_2060 domains are placed with the most similar sequences of the corresponding stand-alone enzymes, bacterial ones in the case of PHP-hydrolase and archaeal ones in the case of the pyruvate formate lyase activating enzyme. The phylogenetic trees produced from these alignments are shown in Figure 2c. The numbers in parentheses show the lengths of regions between the aligned blocks that are not shown. The consensus includes amino acid residues and residue classes that are conserved in 75% of the aligned sequences; the residue classes are as follows: h, hydrophobic; l, alphatic; a, aromatic; s, small; u, tiny; p, polar; b, big; t, residues with high turn-forming propensity. Asterisks show the predicted active site residues; note the replacements in some of the sequences that are predicted to be inactivated versions of the respective enzymes (see text). The alignments were colored using the BOXSHADE program [30]; individual residues conserved in at least 50% of the aligned sequences are in red; residues similar to the conserved ones and groups of conserved similar residues are in blue.
### Table 1

**Interkingdom domain fusions and their probable origins**

| IKF gene (GI number and gene name) and origin of domains | Best 'native' hit | Best 'alien' hit | Protein function | Stand-alone paralog of the alien domain | Comment |
|---------------------------------------------------------|------------------|-----------------|------------------|----------------------------------------|---------|
| Archaea                                                 |                  |                 |                  |                                        |         |
| *Aeropyrum pernix*                                       |                  |                 |                  |                                        |         |
| 5106104_APE2400                                          | 2621953_Mth 1e-27; 282-445; uncharacterized domain conserved among archaea (homolog of the amino-terminal domain of sialic acid synthase) | 2633525 Bs 3e-54; 16-272; hydroxymethyl-pyrimidine phosphate kinase | Hydroxymethyl-pyrimidine phosphate kinase involved in thiamine biosynthesis (additional function?) | None | Pyrococci encode proteins with the same domain organization and closest similarity to *A. pernix*; *M. jannaschii* encodes a protein with the same domain organization but low similarity; *Mt* encodes a HMP-kinase with moderate similarity. |
| Methanococcus jannaschii                                   | 2128140_Mj 1e-19; 2-94; uncharacterized domain | 7270033_At 3e-03; 120-222; A/G2-like stress-related protein | Unknown; possible role in stress response | None | The amino-terminal domain is present in several stand-alone copies in *M. jannaschii*, but otherwise, is seen mostly in bacteria; the possibility of acquisition of a bacterial gene by the Methanococcus lineage is conceivable. |
| Methanobacterium thermoautotrophicum                       | 5103547_Ap 1e-34; 137-26; 5'-formyl-tetrahydrofolate cyclo-ligase | 1651798 Ssp 6e-19; 233-390; GTPase | Membrane-associated 5-formyl-tetrahydrofolate cyclo-ligase(?) exact function unknown | None | In Ssp, the amino-terminal domain is fused to another uncharacterized domain. An ortholog with conserved domain organization is seen in Mycobacterium, but many other bacteria encode stand-alone versions of this domain, which could be the actual sources of horizontal gene transfer. |
| Bacteria                                                 |                  |                 |                  |                                        |         |
| *Aquifex aeolicus*                                        | 2633526_Bs 5e-65; 325-795; c-di-GMP phosphodiesterase | 2650176_Af 6e-19; 116-279; PAS/PAC domain | GTPase, possible role in signal transduction | None | |
| Table 1 (continued) | IKF gene (GI number and gene name) and origin of domains | Best ‘native’ hit (E-value, amino acid residue range, species)/domain function | Best ‘alien’ hit (E-value, amino acid residue range, species)/domain function | Protein function | Stand-alone paralog of the alien domain | Comment |
|---|---|---|---|---|---|---|
| **Bacillus subtilis** | 2632283_yaaH, 1945087_ydhD | 4980914_Tm 1e-06 2-92; LysM repeat domain | 399377_Rn 2e-11 221-402; chitinase | Chitinase | 2635915 | 8. subtilis encodes two paralogous proteins with the same domain architecture |
| **Bacillus subtilis** | 2633242_yhcR | 64819_Dr 1e-64 584-1068; 5’-nucleotidase; 1175987_ECR100; 2e-09 377-521; thermonuclease | 2622704_Mth 0.008 151-257; nucleic acid-binding domain (OB-fold) | Nuclease-nucleotidase (probable repair enzyme) | None |
| **Chlamyphyla pneumoniae** | 4377077_ Cp0769 | 730965_Bs; e-148; 1-727; DNA topoisomerase | 3581917_Sp; 0.003; 7-125; SAM-dependent methyl-transferase | Methyl-transferase/ pyro-phosphatase (metabolic enzyme of an unknown pathway?) | None | Other than in chlamydiae, the SWI domain is seen in eukaryotic chromatin-associated proteins, leading to the suggestion that chlamydial topoisomerase is involved in chromosome condensation |
| **Deinococcus radiodurans** | 6459294_DR1533 | 7248325_Sco; 0.001; 171-265; Mrca family endonuclease | 6754878_Mm; 9e-28; 4-148; G9a domain (DNA-binding) | DNase | None | The G9a domain is not detectable in other prokaryotes. In eukaryotes, this domain so far has been found only as part of multidomain nuclear proteins, including transcription factors |
| **Escherichia coli** | 1787179_b0947 | 94933_Ppu; 3e-10; 287-367; ferredoxin | 3747107_Rn; 3e-32; 4-261; uncharacterized domain (thiol oxidoreductase?) | Oxidoreductase | None | The eukaryotic domain is present (as a partial sequence) also in the beta-proteobacterium Vogesella. This domain contains a conserved pair of cysteines, which together with the ferredoxin fusion, may suggest a thiol oxidoreductase activity. Most of the eukaryotic proteins containing this domain appear to be mitochondrial, suggesting the possibility of an alternative evolutionary scenario |
| **Escherichia coli** | 1787678_b1410 | 487713_Sli; 3e-05; 4e-52; | 5459012_Pab; 1e-17; 33-274; lyso-phospholipase | Methyl-transferase/ Lipase (exact function unclear) | None |
| **Archaeal-archaeal** | 1787679_ymbD | 1591375_Mj; 4e-04; 50-218; membrane-associated acid phosphatase | 7160233_Sp; 1e-06; 346-415; tyrosine phosphatase | Membrane-associated bifunctional phosphatase | None | An unusual case of fusion between an apparently archaeal and a typical eukaryotic domain in a bacterium |

http://genomebiology.com/2000/1/6/research/0013.5
Table 1 (continued)

| IKF gene (GI number and gene name) and origin of domains | Best 'native' hit | Best 'alien' hit | Protein function | Stand-alone paralog of the alien domain | Comment |
|----------------------------------------------------------|-------------------|------------------|------------------|----------------------------------------|---------|
| 1788589_b2255 | 5763950_Sco; 4e-35 | 3860247_At; 1e-55 | Bifunctional enzyme; exact function unclear | None | |
| Bacterial-eukaryotic | methionyl-tRNA formyl-transferase | 1-259; 318-652; dehydratase | | | |
| 1788938_yfiQ | 929735_Nsp; 8e-32; 637-874 | acetyl-Co synthetase/ acetyl-transferase | | | |
| bacterial-Archaeal-eukaryotic | 2649370_Af; 4e-85; 6-689 | acetyl-Co synthetase | | | |
| 1788939_1788589 | 4151109_Tbr; 6e-04; 6-167 | Adenyl cyclase/ ATPase; probable transcription regulator | | | |
| M. tuberculosis encodes three | | | | | M. tuberculosis |
| paralogous proteins that consist | | | | | of three domains, the |
| of three domains, the eukaryotic- | | | | | eukaryotic |
| type adenylate cyclase, AP (apoptotic) | | | | | ATPase and DNA- |
| binding response regulator, and | | | | | two stand-alone versions of |
| adenylate cyclase, which show the | | | | | cyclase domain of the |
| closest similarity to the cyclase | | | | | multidomain |
| domain of the multidomain |
| 1314025_Rv0886 | 120037_Tt; 1e-11; 2e-79; ferredoxin | 178213_Hs; 4e-65; 93-543; ferredoxin reductase | Ferredoxin/ ferredoxin reductase | 2076681 | |
| Bacterial-eukaryotic | | 279520_Dd; 7e-07; 30-105 | cAMP-dependent acetyl-transferase() | (M. leprae) | |
| 3261732_Rv0998 | 2661695_Sco; 3e-13; 148-328; acetyl-transferase | 2645721_Mm; 6-26; 456-972; very-long-chain acyl-CoA synthetase | Bifunctional enzyme of poly (3-hydroxy-butyrate) synthesis | 1929080 | |
| Bacterial-eukaryotic | | 264-251; 8e-27; trehalose-6-phosphate synthase | | | |
| 2326726_Rv1683 | 421331_Cvi; 1e-24; 2e-23-359; poly (3-hydroxy- butyrate) synthase | 389271_At; 8e-27; 264-521; trehalose-6-phosphate synthase | Polyfunctional enzyme of trehalose metabolism | 2661651 | In this protein, the domain of apparent eukaryotic origin is flanked by bacterial domains from both sides |
| Bacterial-eukaryotic | | | | | |
| 1403447_Rv2006 | 675238_Sco; 2e-27; 23-240; phosophatase; 644875_Sco; 0.0; 534-1320; trehalose hydrolyase | | | | |
| Bacterial-eukaryotic | | | | | |
| 2896788_Rv2051c | 117648_Ec; 1e-16; 94-514; apolipoprotein N-acyltransferase | 3073773_Mm; 4e-31; 588-829; dolichol-phosphate-mannose synthase | Polynfuncional enzyme of lipid metabolism | 2337823 | (M. leprae); 6468712 (Streptomyces coelicolor) |
| Bacterial-eukaryotic | 6225563_Scy; 7e-16; 36-253; phosphoserine phorphatase | 1098605_Cnu; 5e-22; | | | |
| 2791523_Rv2483c | 6225563_Scy; 7e-16; 36-253; phosphoserine phorphatase | 1098605_Cnu; 5e-22; | | | |
| Bacterial-eukaryotic | 2633801_Bs; 3e-19; 89-208; molybdopterin synthase large subunit (MoaE) | 4538974_At; 3e-06; 2-82; molybdopterin synthase small subunit (MoaD) | Molybdopterin synthase | 2076687 | The same domain organization is seen in D. radiodurans, but in this case, both components appear to be of bacterial origin |
| 2894233_Rv3323c | 2e-9; 89-208; molybdopterin synthase large subunit (MoaE) | | | | |
| Bacterial-eukaryotic | 2633801_Bs; 3e-19; 89-208; molybdopterin synthase large subunit (MoaE) | | | | |
| 2894233_Rv3323c | 2e-9; 89-208; molybdopterin synthase large subunit (MoaE) | | | | |
| Bacterial-eukaryotic | 2633801_Bs; 3e-19; 89-208; molybdopterin synthase large subunit (MoaE) | | | | |
| 2894233_Rv3323c | 2e-9; 89-208; molybdopterin synthase large subunit (MoaE) | | | | |
| IKF gene | Best ‘native’ hit | Best ‘alien’ hit | Protein function | Stand-alone paralog of the alien domain | Comment |
|----------|------------------|-----------------|------------------|----------------------------------------|---------|
| (GI number and gene name) and origin of domains | (E-value, amino acid residue range, species)/domain function | (E-value, amino acid residue range, species)/domain function | | | |
| 2960152_ | 4753872_ Sco; 56-428; transmembrane efflux protein | 466119_ Ce; 7e-20; 549-964; cAMP-binding domain-phosphoesterase | cAMP-regulated efflux pump(?) | 2501688 | M. tuberculosis encodes two strongly similar paralogs with the same domain architecture |
| Rv3728, | 4731342_ Si; 3e-14; | 1591330_ Mj; 3e-58; molybdenum cofactor biosynthesis protein MoaA (Fe-S oxidoreductase) | Bifunctional enzyme of molybdenum cofactor biosynthesis | 1806159 | The amino-terminal domain stand-alone paralog is more similar to archaean homologs than to the stand-alone paralog, but nevertheless, the latter appears to be of archaean origin |
| 7477351_ | 510-776; C5-O-methyl-Transferase (mitomycin biosynthesis) | 7304009_ Dm; 2e-12; 198-384; peptidoglycan recognition protein | Secreted protein | 7649504 (S. coelicolor) | The stand-alone version of the eukaryotic domain is present only in Streptomyces |
| Rv3239c | 404-494; major secreted protein | | | | |
| Bacterial-archaean | | | | | |
| 3261806_ | 40487_ Cg; 3e-12; 404-494; major secreted protein | 320868_ Sc; 2e-13; 290-488; uridine kinase | 320868_ Sc; 2e-13; 290-488; uridine kinase | None | A co-linear ortholog is present in Thermotoga |
| Rv3811 | | | | | |
| Bacterial-eukaryotic | | | | | |
| Treponema pallidum | 7225946_ Nm; 9e-04; 10-154; threonyl-tRNA synthetase (TGS and H3H domains) | | | | |
| 3322964_ | 7725946_ Nm; 9e-04; 10-154; | 3210401_ Sp; 2e-11; 288-475; uridine kinase | | | |
| TP0667 | 68516_ Bs; 3e-07; 11-200; | 3210401_ Sp; 2e-11; 288-475; uridine kinase | | | |
| Bacterial-eukaryotic | 11-200; threonyl-tRNA synthetase (TGS and H3H domains) | | | | |
| Thermotoga maritima | 68516_ Bs; 3e-07; 11-200; | 3210401_ Sp; 2e-11; 288-475; uridine kinase | | | |
| 4981276_ | 577625_ Hs; 1e-39; | 3210401_ Sp; 2e-11; 288-475; uridine kinase | | | |
| TM0751 | 10-154; | | | | |
| Bacterial-eukaryotic | | | | | |
| Eukaryotes | 586134_ Bt; 9e-10; tubulin-tyrosine ligase | 7450047_ Aa; 8e-09; acid phosphatase (SurE) | Bifunctional signal-transduction protein | 5249 (Yarrowia lipolytica) | SurE homologs are not detectable in eukaryotes other than yeasts |
| Saccharomyces cerevisiae | | | | | |
| 536367_ | 586134_ Bt; 9e-10; tubulin-tyrosine ligase | 7450047_ Aa; 8e-09; acid phosphatase (SurE) | Bifunctional signal-transduction protein | 5249 (Yarrowia lipolytica) | SurE homologs are not detectable in eukaryotes other than yeasts |
| Ybr094w | | | | | |
| Eukaryotic/ | 577625_ Hs; 1e-39; | 3238246_ Ct; 5e-27; | Bifunctional biotin-protein ligase | None | An ortholog with an identical domain architecture is present in S. pombe |
| Bacterial-archaean | 10-154; | | | | |
| 1431219_ | 477096_ Gg; 1e-39; | 1653075_ Sp; 7e-17; | heat shock transcription factor | None | An ortholog with an identical domain architecture is present in S. pombe (3327019) |
| YDL141w | 477096_ Gg; 1e-39; | 1653075_ Sp; 7e-17; | heat shock transcription factor | None | An ortholog with an identical domain architecture is present in S. pombe (3327019) |
| Eukaryotic-bacterial | 78-216; heat shock transcription factor domain | 375-503; CheY domain | | | |
| 458922_ | 78-216; heat shock transcription factor domain | 375-503; CheY domain | | | |
| YHR206W | 78-216; heat shock transcription factor domain | 375-503; CheY domain | | | |
| Eukaryotic-bacterial | | | | | |
| 486539_ | 2983676_ Aa; 114616_ At; 3e-34; urophosphorirn III methylase | 2983676_ Aa; 1e-04; 22-188; precorrin-2 oxidase | Siroheme synthase | 2330809 (S. pombe) | S. pombe also encodes a co-linear ortholog (3581882); apparent displacement of the bacterial precorrin-2 oxidase by a distinct Rossmann fold domain |
| YKR069w | 2983676_ Aa; 114616_ At; 3e-34; urophosphorirn III methylase | 2983676_ Aa; 1e-04; 22-188; precorrin-2 oxidase | Siroheme synthase | 2330809 (S. pombe) | S. pombe also encodes a co-linear ortholog (3581882); apparent displacement of the bacterial precorrin-2 oxidase by a distinct Rossmann fold domain |
| Eukaryotic-bacterial | | | | | |
| 1302305_ | 4938476_ At; 5e-65; 324-861; 7,8-dihydro-6-hydroxymethylpterin-pyro-phosphokinase+Dihydro-pteroate synthase | 3212189_ Hi; 5e-05; 62-148; 187-297; dihydro-neopterin aldolase | Multifunctional enzyme of folate biosynthesis | None | Co-linear orthologs in S. pombe (7490442) and Pneumocystis carinii (283062) |
evolutionarily mobile domains (Figure 3). Generally, the small number of identified IKFs compared to the total number of inferred horizontal transfer events and the total number of domain fusions could be compatible with a random model of domain fusion subsequent to lateral gene transfer.

However, the distribution of IKFs among genomes is distinctly non-random, suggesting that such a simple model may be incorrect. Specifically, 12 IKFs were detected in *Mycobacterium tuberculosis* and 10 were found in the yeast *Saccharomyces cerevisiae*, but only a small number or none was identified in each of the other bacterial and archaean genomes (Figure 2, Table 1). The excess of IKFs in *Mycobacterium* is particularly notable, given that the fraction of genes horizontally transferred from archaean and eukaryotes in the mycobacterial genome is only slightly greater than that in most of the other bacteria, and considerably lower than that in the hyperthermophilic bacteria *Aquifex* and *Thermotoga* (K.S. Makarova, L. Aravind and E.V.K., unpublished observations). Similarly, whereas the overall number of domain fusions in *M. tuberculosis* is greater than in most other bacteria, the difference is insufficient to account for the over-representation of IKFs; furthermore, the cyanobacterium *Synechocystis* sp. has an even greater overall number of fusions but does not have any detectable IKFs (Figure 3).

At present, we cannot provide a defendable biological explanation for the comparatively high frequency of IKF in *Mycobacterium*. It is tempting to interpret this trend in terms of adaptation of this bacterium to its relatively recently occupied parasitic niche, but examination of the individual IKF cases does not offer immediate clues in mycobacterial biology. The yeast IKFs clearly represent relatively recent horizontal transfers distinct from the gene influx from the mitochondria following the establishment of endosymbiosis because, under the protocol of IKF detection used here, only those alien domains were identified that have no counterparts in other eukaryotes.

Most of the IKFs are unique, but *B. subtilis*, *M. tuberculosis* and yeast each also encode families of two to three paralogous IKFs, which apparently have evolved by duplication subsequent to the respective fusion events (Table 1). Strikingly, the same IKF, the three-domain uridine kinase, is shared by *Treponema pallidum* and *Thermotoga maritima* (Table 1). Given that these two bacteria are not specifically related and that *Borrelia burgdorferi*, the second spirochete whose genome has been sequenced, encodes a typical bacterial uridine kinase, the presence of a common IKF in *Treponema* and *Thermotoga* cannot be realistically attributed

Table 1 (continued)

| IKF gene (GI number and gene name) and origin of domains | Best 'native' hit (E-value, amino acid residue range, species)/domain function | Best 'alien' hit (E-value, amino acid residue range, species)/domain function | Protein function | Stand-alone paralog of the alien domain | Comment |
|--------------------------------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------|-----------------|---------------------------------------|---------|
| 1419887__YOL066c_Eukaryotic-bacterial                  | 7297709__Dm; 2e-72; 42-408; large ribosomal subunit pseudoU synthase          | 5918510__Sco; 2e-10; 436-574; pyrimidine deaminase                         | Bifunctional RNA modification enzyme                                     | 2213559  (S. pombe) | The known bacterial homologs have a two-domain organization; the evolutionary scenario could have included domain rearrangements |
| 1419865__YOL055c__2132251__YPL258c__2132289__YPR121w Eukaryotic-bacterial | 2462827__At; 1e-39; 22-390; phosphomethyl pyrimidine kinase (thiamine biosynthesis) | 1075360__Hi; 6e-24; 342-549; transcriptional activator                    | Transcriptional regulator of thiamine biosynthesis genes (?)              | None | Yeast encodes three strongly similar paralogs with identical domain organization; co-linear orthologs are present in other ascomycetes |
| 1370444__YPL214c_Eukaryotic-archaeal/Bacterial          | 2746079__Br; 1e-27; 9-233; thiamin-phosphate pyro-phospholase                  | 2648451__Af; 9e-27; 251-531; hydroxyethyl-thiazole kinase                  | Bifunctional thiamine biosynthesis enzyme                                  | None | Except for the one from *A. fulgidus*, all highly conserved homologs of the kinase domain of this protein are bacterial; it appears likely that the *A. fulgidus* gene is the result of horizontal transfer |

The following complete genomes were analyzed. Archaea: *Aeropyrum pernix* (Ap); *Archaeoglobus fulgidus* (Af); *Methanococcus jannaschii* (Mj); *Methanobacterium thermoautotrophicum* (Mth); *Pyrococcus horikoshi* (Ph); Bacteria: *Aquifex aeolicus* (Aa); *Borrelia burgdorferi* (Bb); *Bacillus subtilis* (Bs); *Chlamydia pneumoniae* (Cp); *Deinococcus radiodurans* (Dr); *Escherichia coli* (Ee); *Haemophilus influenzae* (Hi); *Helicobacter pylori* (Hp); *Mycobacterium tuberculosis* (Mt); *Mycoplasma pneumoniae* (Mp); *Rickettsia prowazekii* (Rp); *Synechocystis* sp (Ssp); *Thermotoga maritima* (Tm); *Treponema pallidum* (Tp). No IKFs were detected in the genomes that are not shown in the table. Additional species name abbreviations: *Ae*, *Arabidopsis thaliana*; *Ax*, *Acytobacter xylinus*; *Bn*, *Brassica napus*; *Ce*, *Caenorhabditis elegans*; *Cvi*, *Chromatium vinosum*; *Gg*, *Gallus gallus*; *Hs*, *Homo sapiens*; *Mm*, *Mus musculus*; *Rn*, *Rattus norvegicus*; *Sco*, *Streptomyces coelicolor*; *Sl*, *Streptomyces lavendulae*. 
Figure 2 (and following page)
Examples of phylogenetic trees supporting the contribution of interkingdom horizontal gene transfer to the emergence of interkingdom domain fusions. The names of proteins from different primary kingdoms are color-coded: black, bacterial; pink, archaeal; green, eukaryotic; the domains involved in the apparent IKF are shown in red. Red circles show nodes with bootstrap support >70%, and yellow circles show nodes with 50-70% support. The bar unit corresponds to 0.1 substitutions per site (10 PAM). (a) IKF: Rv1683 (gi|7476858) from *M. tuberculosis*. Fusion of a bacterial poly(3-hydroxy-butyrate) (PHB) synthase and eukaryotic very long chain acyl-CoA synthetase. Note the absence of eukaryotic homologs in the PHB synthase tree and of bacterial homologs other than the two from *M. leprae* in the acyl-CoA synthetase tree. (b) IKF: yeast YOL066c (gi|6324506). Fusion of a eukaryotic pseudouridylate synthetase with a bacterial pyrimidine deaminase. Note the absence of eukaryotic homologs, other than that from *S. pombe*, in the pyrimidine deaminase tree. (c) IKF: aq 138 (gi|2984285) from *Aquifex aeolicus*. This protein is a fusion of a PHP superfamily hydrolase of apparent bacterial origin and a pyruvate formate-lyase activating enzyme of archaeal origin. (d) IKF: yeast YOL055c (gi|1419865), YPL258c (gi|2132251) and YPR121w (gi|2132289) from *S. cerevisiae*. Fusion of a eukaryotic phosphomethylpyrimidine kinase and a bacterial transcriptional activator. Species abbreviations: Bac.meg., *Bacillus megaterium*; Chr.vin., *Chromatium vinosum*; Thi.vi., *Thiocyts violaceae*; Am.med., *Amycolatopsis mediterranei*; Coch.het., *Cochliobolus heterostrophus*; Dme, *Drosophila melanogaster*; Cel, *Caenorhabditis elegans*; Mus, *Mus musculus*; Spo, *Schizosaccharomyces pombe*; Ath, *Arabidopsis thaliana*; Strep.co., *Streptomyces coelicolor*; The.nea., *Thermotoga neapolitana*; Bac. am, *Bacillus amyloliquefaciens*; Shi.fl., *Shigella flexneri*; Hsa, *Homo sapiens*.

Two evolutionary issues pertaining to IKFs need to be addressed, namely the mechanism(s) of their origin and the selective forces responsible for their preservation. From general considerations, it seems likely that IKFs have
evolved via a two-step process, which involves lateral transfer of the complete gene coding for the IKF’s alien portion, followed by domain fusion. This scenario rests on the assumption that the acquired foreign gene is selectively advantageous, because otherwise it would have been inactivated by mutations before recombination could take place. Under this mechanism, the alien portion of an IKF is likely to be present in the recipient genome also as a stand-alone gene. A clear-cut case of such a duplication of a horizontally transferred domain has been noticed in Chlamydia, whose genomes encode the SWI domain, implicated in chromatin condensation, both as a stand-alone protein and as the carboxy-terminal portion of topoisomerase I [10]. Apart from this case, the IKFs fall into two readily discernible classes, namely those from Mycobacterium and all the rest. M. tuberculosis (the only complete genome of an actinomycete available) possesses considerably more IKFs than any other bacterial or archaeal species (see above), and typically, the alien portions of these proteins show high level of similarity to the homologs from the donor superkingdom (eukaryotes). Most significantly, there is also, with a single exception, a stand-alone counterpart in the mycobacterial genome; in some cases, such a counterpart is seen only in a closely related species, M. leprae, and in one case, it is found in Streptomyces, a distantly related actinomycete (Table 1). In the other genomes, the IKFs are generally less similar to
the apparent donor and, with a few exceptions, stand-alone versions of the alien domains are missing (Table 1). The hypothesis that seems to be most compatible with these observations is that IKFs indeed evolve via a stand-alone, horizontally transferred intermediate, but in the case of ancient IKFs, these intermediates are typically eliminated during evolution, perhaps because their function becomes redundant with the formation of the IKF. The IKFs identified in actinomycetes appear to result from relatively recent gene fusion events so that the original, stand-alone transferred genes are still present in the genome.

The IKFs include a variety of protein functions. Only some of these are well understood such as, for example, those of the bifunctional nucleotide and coenzyme metabolism enzymes that are particularly abundant in yeast (Table 1). In other cases, the function of an IKF-encoded protein could be predicted only tentatively on the basis of the functions of its constituent domains (Table 1). The selective advantage of the formation of multidomain proteins, at least as far as enzymes are involved, lies in the possibility of effective coupling of the reactions catalyzed by the different domains [16]; this may be generalized also for functional coordination of non-enzymatic domains. Fusion may result in the addition of a regulatory function to an enzymatic one. For example, it appears most likely that the RNA-binding TGS domain [24] in the uridine kinases of Treponema pallidum and Thermotoga maritima is involved in autoregulation of translation. The unusual aspect of the IKFs appears to be the compatibility of evolutionarily distant domains.

Examination of the phyletic distribution of the multidomain architectures of IKFs may help in pinpointing the evolutionary stage at which the fusion (but not necessarily the preceding horizontal gene transfer) has occurred. For example, the fusion of the SWI domain with topoisomerase belongs after the radiation of Chlamydia from other bacterial lineages, but before the radiation of Chlamydia pneumoniae and Chlamydia trachomatis.
trachomatis (Table 1). The majority of IKFs detected in the yeast *S. cerevisiae* are also present in *Schizosaccharomyces pombe* and/or other ascomycetes (Table 1, and data not shown), but not in any other eukaryotes, and accordingly, they should have evolved at a relatively early stage of fungal evolution, but not before the fungal clade diverged from the rest of the eukaryotic crown group.

Finally, it should be noted that formation of some of the IKFs might have required more complex rearrangements of the contributing proteins than simple domain fusion. Figure 4 shows the domain architectures of proteins that contribute domains to two IKFs. In each case, a simple fusion between genes encoding the respective individual domains is insufficient to explain the emergence of the IKF. For example, the uridine kinase example mentioned above (Figure 4a) should have involved isolation of the TGS-HxxH domains of threonyl-tRNA synthetase before or concomitantly with their fusion with the uridine kinase. The specific molecular mechanism could have involved selective duplication of the upstream portion of the threonyl-tRNA synthetase gene. Similarly, the sialic acid synthase homologous domain, which is fused to hydroxymethylpyrimidine phosphate kinase in *A. pernix* and pyrococci, appears to have been derived from two-domain proteins that additionally contain a helix-turn-helix DNA-binding domain (Figure 4b). These hypotheses of a complex mechanism of gene fusion involved in the emergence of IKFs are based on a limited sample of sequenced genomes. An alternative possibility is that, before the postulated horizontal transfer event, the recipient domain(s) has been encoded by a stand-alone gene; such genes that do not contain the fused alien domain may yet be discovered in newly sequenced genomes. In fact, a stand-alone version of the sialic acid synthase homologous domain is seen in *Methanobacterium*, although it is considerably less similar to the IKF than the version fused to the HTH domain (Figure 4b).

The identification of IKFs underscores the complexity of the evolutionary process as revealed by comparison of multiple genomes. In and by itself, this phenomenon may not have a unique biological significance, but it reveals the overlap between two major evolutionary trends, horizontal gene transfer and protein domain rearrangement, and shows that domains, rather than entire proteins (genes), should be considered fundamental units of genetic material exchange.

**Materials and methods**

Protein sequences encoded in 21 complete genomes of archaea, bacteria and the yeast *Saccharomyces cerevisiae* were extracted from the Genome division of the Entrez retrieval system [25]. Each protein encoded in these genomes was used as the query in a comparison against the non-redundant protein sequence database (National Center for Biotechnology Information, NIH, Bethesda, USA) using the BLASTP program [26]. For each query, the set of local similarities detected by BLASTP was automatically (using a Perl script written for this purpose) screened for putative IKFs, that is situations in which the query did not have full-size homologs outside its immediate taxonomic group (for example, the Proteobacteria for *Escherichia coli*) and in which different regions of the query showed the greatest similarity to proteins from different primary kingdoms. The pseudocode for the script follows:

Let $H$ be \{ $h_1, h_2, \ldots, h_N$ \} - hits for the query $Q$

$Lq$ - query length

$TSq$ - query superkingdom

$TFq$ - query family

for each $p < Lq$ {

\[ No \leftarrow 0 \]

\[ TSbestL \leftarrow "" \]

\[ TSbestR \leftarrow "" \]

for each $h$ in $H$ by decreasing score {

\[ TSh \leftarrow hit superkingdom \]

\[ TFh \leftarrow hit family \]

if($TFh$ == $TFq$)\{ next $h$ \}

\[ Ph(p) \leftarrow position of h relative to p \]

if($Ph(p)$ == overlap) {

if($TSh$ != $TSq$)\{ next $h$ \}

\[ No \leftarrow No + 1 \]

if($No >= maximum allowed No)$\{ exit \}

}\elseif($Ph(p)$ == left) {

if($TSbestL$ not empty)\{ next $h$ \}

\[ TSbestL \leftarrow TSh \]

}\elseif($Ph(p)$ == right) {

if($TSbestR$ not empty)\{ next $h$ \}

\[ TSbestR \leftarrow TSh \]

}\}

if($TSbestL$ != $TSbestR$) {

report $Q$ as a candidate to IKF

exit

}

exit

The script itself is available as an additional data file with the online version of this paper. The candidate IKF cases were further examined to detect situations where one or more distinct regions of the query could be classified as ‘native’ or ‘alien’ either on the basis of the lack of close homologs from the respective primary kingdom or using phylogenetic analysis. Multiple sequence alignments were generated using the ClustalW program [27], and when necessary, manually corrected to ensure the proper alignment of conserved motifs typical of the respective domains. Phylogenetic trees were constructed using the PROTDIST and FITCH programs of the PHYLIP package [28]. Trees were made separately for each domain of a putative IKF, and its mixed ancestry was considered confirmed if the affinities of the domains with different primary kingdoms were supported by bootstrap values of at
least 50%. Additional iterative database searches were performed using the PSI-BLAST program [26,29] in order to predict functions of the individual domains of the identified IKFs in cases when these were not immediately clear.

Additional data
The following additional data are included with the online version of this paper: the Perl script used to screen local similarities for putative IKFs.

References
1. Doolittle WF: Lateral genomics. Trends Cell Biol 1999, 9:MS-MB.
2. Doolittle WF: Phylogenetic classification and the universal tree. Science 1999, 284:2124-2129.
3. Koonin EV, Mushegian AR, Galperin MY, Walker DR: Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol Microbial 1997, 25:619-637.
4. Ochman H, Lawrence JG, Groisman EA: Lateral gene transfer and the nature of bacterial innovation. Nature 2000, 405:299-304.
5. Gray MW: Evolution of organelar genomes. Curr Opin Genet Dev 1999, 9:678-687.
6. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000, 28:33-36.
7. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. Science 1997, 278:631-637.
8. Aravind L, Tatusov L, Wolf YI, Walker DR, Koonin EV: Evidence for massive gene exchange between archaea and bacterial hyperthermophiles. Trends Genet 1998, 14:442-444.
9. Nelson KE, Claydon RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, et al.: Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 1999, 399:323-329.
10. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, et al.: Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. Science 1998, 282:754-759.
11. Subramanian G, Koonin EV, Aravind L: Comparative genome analysis of pathogenic spirochetes - Borrelia burgdorferi and Treponema pallidum. Infect Immun 2000, 68:1633-1648.
12. Wolf YI, Aravind L, Koonin EV: Rickettsiaceae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends Genet 1999, 15:173-175.
13. Doolittle RF: The multiplicity of domains in proteins. Annu Rev Biochem 1995, 64:287-314.
14. Enright AJ, Iliopoulos I, Kyprides NC, Ouzounis CA: Protein interaction maps for complete genomes based on gene fusion events. Nature 1999, 402:86-90.
15. Galperin MY, Koonin EV: Who is your neighbor: new computational approaches in functional genomics. Nat Biotechnol 2000, 18:609-613.
16. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: Detecting protein function and protein-protein interactions from genome sequences. Science 1999, 285:751-753.
17. Marcotte EM, Pellegrini M., Thompson MJ, Yeates TO, Eisenberg D: A combined algorithm for genome-wide prediction of protein function. Nature 1999, 402:83-86.
18. Aravind L, Koonin EV: DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res 1999, 27:4658-4670.
19. Grebe TW, Stock JB: The histidine protein kinase superfamily. Adv Micro Physiol 1999, 41:139-227.
20. Grebe TW Jr, Reiter J.: The bacterial phosphotransferase system: new frontiers 30 years later. Mol Microbiol 1994, 13:755-764.
21. Koonin EV, Aravind L, Kondrashov AS: The impact of comparative genomics on our understanding of evolution. Cell 2000, 101:573-576.
22. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 1999, 9:608-628.
23. Wolf YI, Aravind L, Grishin NV, Koonin EV: Evolution of aminoacyl-tRNA synthetases - analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 1999, 9:689-710.
24. National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html]
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25:3389-3402.
26. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673-4680.
27. Altschul SF, Koonin EV: PSI-BLAST - a tool for making discoveries in sequence databases. Trends Biochem Sci 1998, 23:444-447.
28. http://www.ch.embnet.org/software/BOX_form.html