Integration of Horizontally Transferred Genes into Regulatory Interaction Networks Takes Many Million Years

Martin J. Lercher and Csaba Pál†‡*

*Department of Computer Science, Heinrich-Heine-University, Düsseldorf, Germany; †Centre for Computational and Systems Biology, Microsoft Research, University of Trento, Trento, Italy; and ‡Systems Biology Unit, Institute of Biochemistry, Biological Research Center, Szeged, Hungary

Adaptation of bacteria to new or changing environments is often associated with the uptake of foreign genes through horizontal gene transfer. However, it has remained unclear how (and how fast) new genes are integrated into their host's cellular networks. Combining the regulatory and protein interaction networks of *Escherichia coli* with comparative genomics tools, we provide the first systematic analysis of this issue. Genes transferred recently have fewer interaction partners compared to nontransferred genes in both regulatory and protein interaction networks. Thus, horizontally transferred genes involved in complex regulatory and protein–protein interactions are rarely favored by selection. Only few protein–protein interactions are gained after the initial integration of genes following the transfer event. In contrast, transferred genes are gradually integrated into the regulatory network of their host over evolutionary time. During adaptation to the host cellular environment, horizontally transferred genes recruit existing transcription factors of the host, reflected in the fast evolutionary rates of the *cis*-regulatory regions of transferred genes. Further, genes resulting from increasingly ancient transfer events show increasing numbers of transcriptional regulators as well as improved coregulation with interacting proteins. Fine-tuned integration of horizontally transferred genes into the regulatory network spans more than 8–22 million years and encompasses accelerated evolution of regulatory regions, stabilization of protein–protein interactions, and changes in codon usage.

Introduction

Horizontal gene transfer (HGT) is widely recognized as a major contributor to evolutionary innovations in bacterial lineages (Ochman et al. 2000; Koonin et al. 2001; Gogarten et al. 2002; Jain et al. 2002; Daubin et al. 2003; Koonin 2003; Nakamura et al. 2004; Lerat et al. 2005). Gene transfer can be viewed as a process comprised of 3 consecutive steps: 1) the physical transfer of DNA into a foreign cell, followed by the integration of the DNA into the genetic repertoire of the individual host cell; 2) the spread of this variant through the bacterial population (fixation)—this will mostly be due to selective advantages provided by the transferred DNA and is hence necessarily associated with a preliminary functional integration of the new gene into the cellular environment; and 3) the fine-tuning of biochemical interactions, caused by selective pressure to optimize the functional integration and resource usage of the new gene. The first step is independent of the functional characteristics of the transferred genes and its details are not considered below.

It is obvious that no horizontally transferred gene can work in isolation: appropriate transcription, translation, protein folding, and multiunit complex formation require various molecular signaling mechanisms. These signals, encoded in the foreign gene sequence and its associated *cis*-regulatory regions, must be recognized by host proteins and various cellular machineries; for example, initiation and rate of transcription/translation depend on the recognition of promoters, transcription factor–binding sites, and ribosomal binding sites. In a similar vein, appropriate subcellular localization demands recognizable signal peptides and proper multiunit complex formation requires precise coregulation to ensure correct stochiometry of subunits.

These various layers of cellular recognition evolve. Thus, appropriate interaction of transferred genes with the host cell is expected to diminish with increasing phylogenetic distance between host and donor organisms. Accordingly, the likelihood of fixation of a transferred gene diminishes, whereas the necessity to fine-tune established interactions increases. Although some of the above problems can partly be relieved through the cotransfer of interaction partners, there are limits on the length of simultaneously transferred DNA; these limits prevent the successful cointegration of members of large, heavily interacting functional modules in a single step.

Accordingly, the "complexity hypothesis" posits that genes involved in complex cellular subsystems are less likely to undergo HGT (Rivera et al. 1998; Jain et al. 1999, 2002). Whereas the differential rates of HGTs across functional classes are consistent with this hypothesis (Rivera et al. 1998; Jain et al. 1999), the theory remained hotly debated (Brochier et al. 2000; Nesbo et al. 2001). A recent study (Wellner et al. 2007) concluded that horizontally transferred genes are indeed less involved in protein–protein interactions. Based on the comparison of 2 different methods for identifying horizontally transferred genes, it was suggested that transferred genes integrate slowly into existing protein–protein interaction networks (Wellner et al. 2007).

Given the potentially harmful consequences of unregulated expression of foreign genes (including transposable elements and viruses), several safeguard mechanisms have evolved to detect and silence horizontally transferred genes (Navarre et al. 2006; Dorman 2007). For example, the heat-stable nucleoid-structuring system—which is widespread across enterobacterial species including *Escherichia coli*—has a key role in selectively silencing the transcription of GC-poor genes, including many genes derived from horizontal transfers (Dorman 2007). This repression mechanism can later be relieved by the formation of new regulatory

Key words: DNA regulatory network, horizontal gene transfer, protein–protein interactions.

E-mail: lercher@cs.uni-duesseldorf.de.

Mol. Biol. Evol. 25(3):559–567. 2008
doi:10.1093/molbev/msm283
Advance Access publication December 24, 2007

interactions with host transcription factors (Dorman 2007), as well as through the slow "amelioration" of GC content to that found in "native" host genes (Lawrence and Ochman 1998). Low levels of gene expression are likely to be especially problematic for genes involved in the formation of multiunit protein complexes, as low dosage impedes complex formation (Deutschbauer et al. 2005), and imbalance of complex subunits can lead to harmful protein aggregation (Papp et al. 2003).

Here, we focus on the coevolution of bacterial protein–protein interaction and regulatory networks when "perturbed" by gene transfer events. We focus on the integration of novel genes and are not concerned with xenologous replacements. Employing a method that allows the dating of horizontal transfer events, we systematically contrast nontransferred genes and transferred genes of different age groups. Thereby, we examine 2 questions: can novel genes easily establish physical and regulatory interactions with genes of their new host? And how are novel genes incorporated into the regulatory systems of their host over evolutionary time?

Materials and Methods

Identification of Orthologs

From the National Center for Biotechnology Information, we downloaded the protein sequences of 31 fully sequenced bacteria that are derived from the last common ancestor of *E. coli* and *Shewanella oneidensis*. We performed reciprocal BlastP searches of all proteome pairs (Altschul et al. 1997). Orthologs were selected based on reciprocal best Blast hits using an E value cutoff of 10^{-40} (other cutoff values led to very similar results, see supplementary tables 1–5 [Supplementary Material online]). Orthologous clusters of genes were defined as groups of sequences (0 or 1 per species), where each sequence had each of the other group members as the best BlastP hit in the respective proteome. If any of the sequences in a cluster had a reciprocal best Blast hit in another proteome, but the latter sequence was not best reciprocal hit to all other sequences already included in the group, the whole group was excluded from further analyses. This requirement of all-against-all reciprocal best hits is very stringent and thus gives good confidence in the inferred orthology. Ortholog identification was first performed for all 31 species to allow phylogenetic reconstruction including 7 outgroup species (see below). To identify orthologous clusters for the analysis of HGTs, we then repeated the protocol for the 24 ingroup species (3 species were later excluded, see below).

Phylogenetic Reconstruction

To reconstruct the phylogenetic relationships among the 31 strains of bacteria, we selected all 114 orthologous clusters that contained one sequence from each strain. Multiple sequence alignments were performed with MUSCLE using default settings (Edgar 2004). Alignments were purged from unreliably aligned positions as well as gaps with Gblocks (Castresana 2000), requiring that flanking regions of blocks were conserved across all strains. The remaining 42,377 positions were concatenated for phylogenetic analysis. We performed a maximum likelihood analysis with phyML (Guindon and Gascuel 2003), employing an empirical substitution model (Jones et al. 1992) and accounting for among-site rate variation with a discrete Γ-model. All but 3 branches of the resulting phylogeny were supported by >95% of 1,000 bootstrap replicates; To be conservative, we restricted all analyses to the 21 species with a fully resolved phylogeny (fig. 1; excluding *Salmonella enterica Cholerasuis*, *Salmonella enterica Paratyphi ATCC9150*, and *Yersinia Pestis KIM* and treating *Shewanella* and the *Vibrio*/*Photobacterium* clade as outgroups).

Inference of HGT

Based on the presence and absence of genes in each of the orthologous clusters of genes across the 21 proteobacterial species (excluding singletons, i.e., genes that have no reciprocal best Blast hit in any other species), we reconstructed the most parsimonious scenarios for gene loss

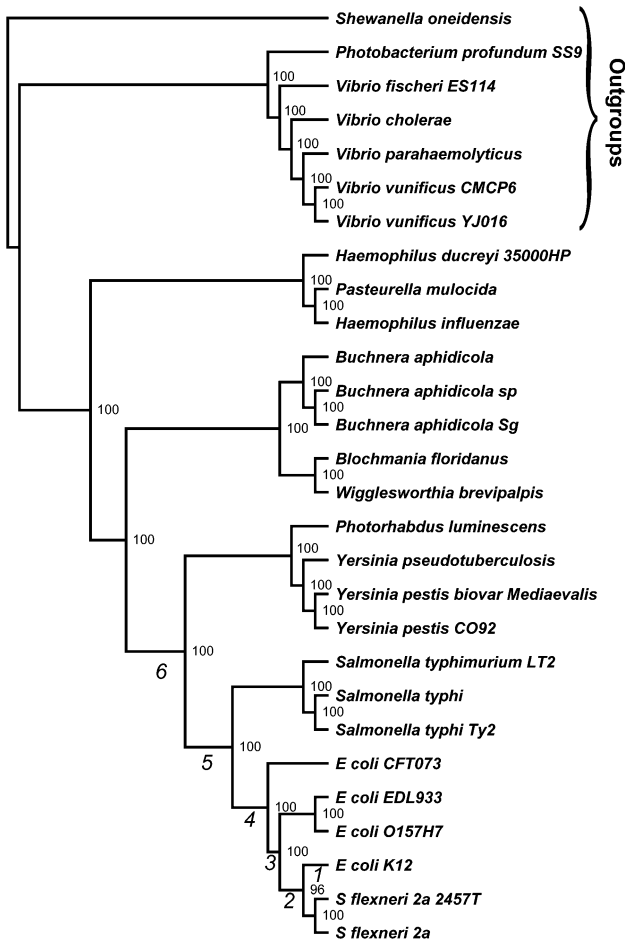

FIG. 1.—Bacterial phylogenetic tree. The tree was derived using maximum likelihood analysis of ubiquitous single-copy genes under an empirical model of amino acid substitutions. Labels to the right of branches give bootstrap support in percent. The branching order agrees with previously published results. Italic labels below branches indicate the relative age of horizontal transfers into the *Escherichia coli* K12 lineage assigned to these branches.

and horizontal transfer events (gene gains) on the rooted phylogeny using generalized parsimony as implemented in PAUP* version 4.0 (see [Pal et al. 2005] and further references therein). All results in the main text are obtained using the DELTRAN algorithm, with relative penalties for HGT and loss of 2:1. This penalty ratio was previously shown to be biologically realistic (Snel et al. 2002). Results obtained for different settings are very similar (supplementary tables 1–5, Supplementary Material online). For further analyses, we considered only orthologous clusters containing a sequence from *E. coli* K12 that had an associated Blattner name. To measure the age of individual transfer events into *E. coli* K12, we used the number of nodes separating *E. coli* K12 from the branch where the event occurred, starting with a value 1 for the terminal *E. coli* K12 branch itself (fig. 1).

Network and Expression Analyses

The protein interaction network of *E. coli* K12 was reconstructed by merging 2 large-scale pulldown screens (Butland et al. 2005; Arifuzzaman et al. 2006), yielding 16,384 protein interactions. For each protein used as a bait in either one or both studies, we defined connectivity as the total number of different prey proteins that were identified in the 2 studies. We restricted our attention to genes that 1) were systematically screened against nearly all *E. coli* proteins in at least one of these studies and 2) for which reliable information was available on the presence/absence of its orthologs across the phylogenetic tree. The number of protein interactions was calculated for all 1,803 genes that fulfilled the above criteria.

The microarray gene expression data set used for the coexpression analysis was derived from the Stanford Microarray Database (Ball et al. 2005), as compiled and normalized across experiments by Bhardwaj and Lu (2005). The data set includes 52 data points, measured under several biological conditions, including RNA degradation, nutrient limitation, and enzyme inhibition/promotion (Bhardwaj and Lu 2005). We concentrated on protein interactions with 1) appropriate expression data for both genes and 2) evidence that at most one of the partners has been transferred into the *E. coli* K12 lineage, leaving 7,605 interacting gene pairs. As previously (Bhardwaj and Lu 2005), we used Pearson's correlation coefficient (r) as the measure of mRNA coexpression among interacting proteins.

Transcriptional regulatory interactions and information on computationally predicted transcriptional units were downloaded from the updated RegulonDB database (Salgado et al. 2006), which contains experimental evidence on 150 regulators affecting 2,862 genes in *E. coli* K12.

Calculation of Sequence Conservation at Upstream Noncoding Regions

Upstream intergenic DNA regions of 1,838 orthologous gene pairs were extracted from the genome sequences of *E. coli* K12 and *Shigella flexneri 2a*. The data were limited to the intergenic region up to a maximum of 500-bp upstream of the translation start site, regions where almost all bacterial control elements are found (Gralla and Collado-Vides 1996). To avoid statistical nonindependence, only genes next to the promoter region of each transcription units were analyzed further (Salgado et al. 2006). Sequence pairs were aligned with MCALIGN2 (Wang et al. 2006), using an indel frequency model specified for noncoding DNA upstream of genes. To ensure that only homologous noncoding sequences were aligned, only relatively long ($\geq$50 aligned nucleotides) and well-conserved ($\geq$90% sequence identity) alignments were retained for subsequent analyses. Evolutionary distances were calculated using the method of Jukes and Cantor (Nei and Kumar 2000). Other methods give nearly identical results (supplementary table 4, Supplementary Material online). The results are unlikely to be sensitive to specific parameters of the alignment program for 2 reasons. First, as very closely related organisms were analyzed, alignments contained few indels (data not shown). Second, distances were also calculated based on alignments obtained by SIGMA (Siddharthan 2006), a recently developed program for the alignment of noncoding sequences. Results based on the 2 alignment methods show excellent agreement (Pearson's correlation coefficient $r = 0.973$).

Results and Discussion

Identifying Horizontally Transferred Genes

Several methods exist for detecting horizontally transferred genes (Gogarten et al. 2002). They are based on 1) atypical sequence composition, 2) patchy or limited phylogenetic distribution across related species, or 3) incongruence between phylogenetic gene and species trees. Crucially, these methods examine different properties of the genomes, identify different subsets of horizontally transferred genes, and are therefore appropriate for testing different types of hypotheses (Gogarten et al. 2002). A significant difference has recently been pointed out between the identification of xenologous replacements of genes by clear orthologs obtained from other species and the identification of integrated foreign genes (e.g., pathogenicity islands) with no allelic counterparts in the host or in closely related genomes (Daubin and Ochman 2004; Ge et al. 2005). In particular, the reliance of phylogenetic incongruence methods on the existence of orthologs in several related species means that they cannot be applied to genes with limited phylogenetic distributions.

Problems concerning the functional integration of new elements without already present orthologs are likely to differ from those associated with xenologous replacements (Koonin et al. 2001). As our work aims to elucidate the integration of novel genes, we used a method optimized for the detection of the former type of transfer. We mapped gene gain and loss events onto a phylogenetic species tree, using established protocols based on the distribution of genes across species. This approach has the desired effect of ignoring xenologous replacements of genes with (nearly) identical functions and properties. However, our main conclusions do not depend on the choice of method for the identification of transferred genes: Very similar results are obtained when we restrict our analyses to transfers confirmed through complementary methods (see below).

We first assembled stringent orthologous families of genes across 21 fully sequenced proteobacteria species including *E. coli* K12. As our protocol is especially suitable for analyzing gene gain and loss events among closely related species, we included several *E. coli* strains and their relatives. We then established the species phylogeny based on 114 ubiquitous single-copy genes, using maximum likelihood methods. The resulting tree was not only well supported by bootstrap analyses (fig. 1) but also agreed with previous results on subsets of the investigated species (Mirkin et al. 2003; Boussau et al. 2004; Pal et al. 2005). We then used established protocols (Kunin and Ouzounis 2003; Mirkin et al. 2003; Boussau et al. 2004; Pal et al. 2005) for the inference of horizontal transfer events across 2,977 orthologous gene families with members in *E. coli* K12. We identified the most parsimonious scenario for HGTs and gene losses across the tree based on the presence or absence of proteins from each species. The inferred rates of HGT may be overestimates (Zhaxybayeva et al. 2007) as gene loss commonly proceeds through gradual loss of gene activity. However, results remain after excluding potential pseudogenes and genes with no detectable transcripts in the *E. coli* K12 genome (see below).

Several empirical results confirm the reliability of our method. First, consistent with expectations and earlier observations (Lawrence and Ochman 1998), a substantial fraction (14%) of the most recently transferred genes (branches 1 and 2 in fig. 1) are annotated (Keseler et al. 2005) with virus- or transposon-related functions (supplementary fig. 1*A*, Supplementary Material online). Second, for recently acquired genes, our assignments of horizontal transfers are in good agreement with those from complementary approaches based on atypical gene composition (for details, see [Pal et al. 2005] and supplementary fig. 1*B* [Supplementary Material online]). The observed gradual decay of codon usage and GC content irregularities with the age of the transfer event (supplementary fig. 1*C* and *D*, Supplementary Material online) agrees with the previously hypothesized amelioration of compositional biases over evolutionary time (Lawrence and Ochman 1998). Finally, the results of this paper are not sensitive to modifications of the parameters used in the generalized parsimony analysis: All major results reported below have been confirmed using different parameter settings for horizontal transfer identification (supplementary tables 1–5, Supplementary Material online).

It has recently been recognized that proteins of horizontally transferred genes evolve at especially high rates (Hao and Golding 2006). As the initial step of our HGT detection method hinges on identifying orthologs by sequence similarity, one might thus worry that the method confounded rapid evolution with horizontal transfer. This is unlikely to be a problem: The probability of missing orthologs (and thereby falsely inferring transfers) in fast-evolving genes is expected to increase with more stringent similarity cutoffs used for assembling orthologs. However, varying the Blast cutoff has no consistent effect on the difference in the number of protein interactions between transferred and nontransferred genes (supplementary tables 1–5, Supplementary Material online).

Horizontally Transferred Genes Have Few Physical Protein Interactions

Network position in the protein interaction network of *E. coli* K12 was obtained by combining data from 2 large-scale systematic studies (Butland et al. 2005; Arifuzzaman et al. 2006). Remarkably, 13–24% (supplementary table 1, Supplementary Material online) of the genes with protein interactions in our data set are likely to be the result of horizontal gene transfer since the divergence of *E. coli* from the *Salmonella* lineage (estimated to have occurred approximately 100 MYA [Lawrence et al. 1991]). This identifies HGT as a major force in the evolution of protein interaction networks, analogous to findings on biochemical metabolic networks (Pal et al. 2005).

In agreement with the complexity hypothesis and a recent study based on complementary approaches for gene transfer identification (Wellner et al. 2007), we find that hubs of the protein interaction network are rarely transferred: The mean number of horizontal transfers across the tree decreases drastically with increasing numbers of interactions (fig. 2*A*). Genes most susceptible to HGT (those transferred more than once across the phylogenetic tree) have on average 70% fewer interactions compared with genes never transferred (fig. 2*B*).

The negative correlation between HGTs and the number of protein interactions remains when only interactions validated by further experiments are analyzed (supplementary table 1*b*, Supplementary Material online). Transfer events confirmed by incongruence of the phylogenetic gene and species trees (supplementary fig. 2, Supplementary Material online) show similar differences in the number of protein interactions between the 2 classes of genes. Further, the relationship between the number of protein interactions and the number of transfer events across the tree remains after controlling for potential confounding variables (essentiality [Pal et al. 2005], expression level [Taoka et al. 2004], and functional classes [Garcia-Vallve et al. 2000]; [supplementary table 2, Supplementary Material online]). Finally, we found only a weak (though statistically significant) relationship between the number of protein interactions and the rate of protein evolution among nontransferred genes (supplementary fig. 3, Supplementary Material online). Thus, our results are also not caused by confounding rapid evolution with HGT (further confirmed by results from different Blast cutoffs, supplementary tables 1–3 [Supplementary Material online]).

Low connectivity might simply reflect the condition-specific activity of transferred genes. Indeed, in the context of metabolic networks, we have shown previously (Pal et al. 2005) that horizontally transferred genes often confer selective advantages in special environments and are generally located at the network periphery (i.e., function in the uptake or catalysis of external nutrients). However, analysis of growth rates of knockout mutants under 282 environmental conditions reveals no relationship between condition specificity and the number of protein interactions (supplementary fig. 4, Supplementary Material online). Thus, environmental specificity is unlikely to cause the observed pattern.

How does connectivity evolve over evolutionary time? There is at best only a weak relationship between the

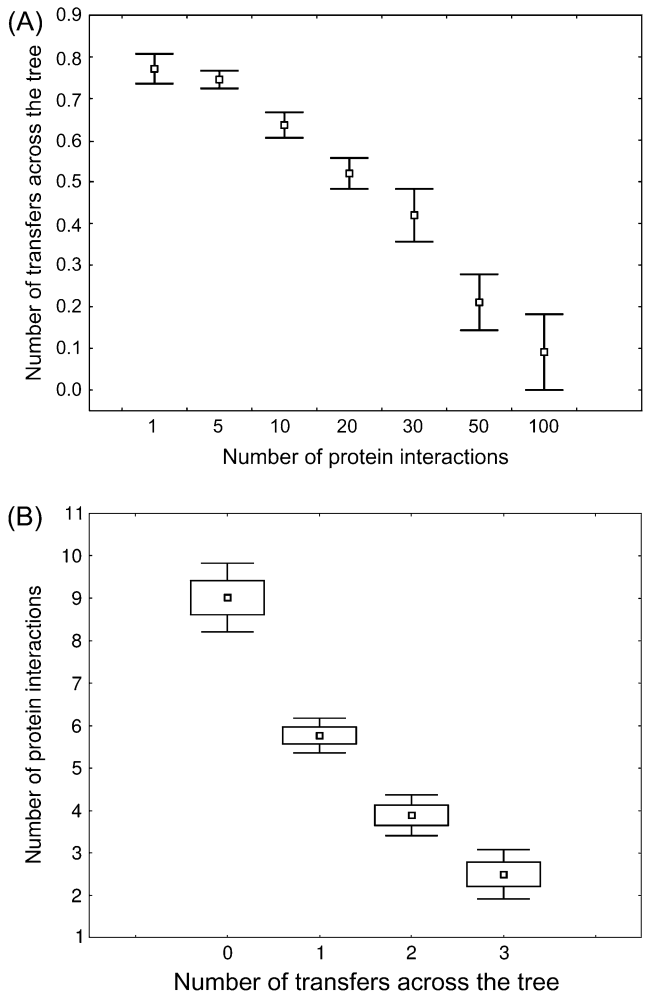

FIG. 2.—HGT and protein interactions. (*A*) Number of gains across the complete phylogenetic tree for groups of proteins with different numbers of protein interactions (analysis of variance [ANOVA], $N = 1{,}803$, $F = 11.42$, degrees of freedom [df] = 6, $P < 10^{-10}$). Central squares are mean, whiskers indicate standard error (SE). (*B*) The numbers of interaction partners for genes with different numbers of inferred horizontal transfers across the complete species tree. (ANOVA, $N = 1{,}803$, $F = 27.23$, df = 3, $P < 10^{-10}$). Central squares indicate mean, upper and lower box margins indicate SE, whiskers indicate 2 × SE. Similar results are obtained when the analysis is restricted to transfer events into the *Escherichia coli* K12 lineage (data not shown).

number of protein interactions and increasing age of the transfer (supplementary table 3, Supplementary Material online). Even proteins resulting from the most ancient transfers into the *E. coli* lineage have much fewer interactions than nontransferred genes (supplementary fig. 5, Supplementary Material online). This implies that physical protein–protein interactions are not often gained (or regained) after horizontal transfers: For a transferred gene to become fixed in the bacterial population, important interaction partners have to be available immediately after transfer. The low connectivity of proteins resulting from transfers across all age groups thus indicates that physical protein–protein interactions indeed impede successful integration into a new host, as predicted by the complexity hypothesis (Jain et al. 1999): Transferred genes are unlikely to perform selectively favorable functions if suitable interaction partners are not already present in their new host.

Only Recently Transferred Genes Have Few Regulatory Interactions

Due to the fast evolution of bacterial transcriptional networks (Madan Babu et al. 2006), transcription factors will often differ significantly between transfer source and target species. Hence, many horizontally transferred genes may initially not be optimally regulated, resulting in low expression levels (Taoka et al. 2004). To test this prediction, we obtained the known transcriptional regulatory network of *E. coli* from RegulonDB (Salgado et al. 2006).

Analogous to our results for protein–protein interactions, we found that genes recently transferred into the *E. coli* K12 lineage are less likely to be under the direct control of known transcriptional regulators compared with nontransferred genes (fig. 3). However, whereas even proteins resulting from ancient transfers have few protein–protein interactions, their transcriptional control approaches that of nontransferred genes (fig. 3). Remarkably, it appears that the average number of transcriptional regulators still increases for proteins up to at least age class 4 (fig. 3), consisting of transfer events that occurred before the diversification of the *E. coli* lineages (fig. 1).

Proteins of recently transferred genes are known to be weakly expressed (Taoka et al. 2004). This observation is consistent with at least 2 theories. First, horizontally transferred genes might represent special classes of genes with low optimal gene expression level. Alternatively, improper interactions with the host cellular machinery might prevent appropriate levels (and timing) of expression. Consistent with the latter idea, the number of transcription factors enhancing the transcription rate of a given gene (positive regulators) strongly increases shortly after horizontal transfer and remains relatively constant afterward. In contrast, the number of transcription factors suppressing the expression level of a given gene (negative regulators) increases more slowly and gradually through evolutionary time (fig. 3*A*). Thus, it appears that transcription is rapidly boosted initially after transfer and then slowly fine-tuned by the addition of negative regulators. The early recruitment of positive regulators may also have a distinct role in relieving the impact of genomic safeguarding mechanisms that silence foreign elements (Dorman 2007).

Accelerated Regulatory Evolution Upstream of Transferred Genes

Regulatory interactions between host transcription factors and horizontally transferred target genes can arise through de novo evolution of transcription factor–binding sites and/or through changes in existing binding sites at the transferred gene. In both scenarios, one would expect accelerated evolution of the transcriptional control regions of recently transferred genes.

To test this prediction, we compared the degree of conservation at homologous upstream DNA regions between *E. coli* K12 and its close relative *S. flexneri*. Genes that were

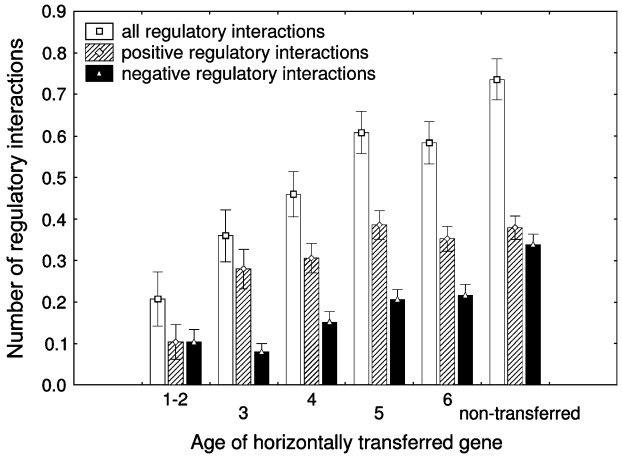

FIG. 3.—Horizontally transferred genes are rarely under the control of known transcription factors; positive interactions are established faster than negative interactions (mean ± SE). The figure shows the average number of transcription factors (or transcriptional regulatory links) per gene versus the age class of the transferred gene. Self-regulatory interactions were excluded. Statistics: all interactions: ANOVA, $N = 2{,}759$, df = 5, $F = 6.37$, $P < 5 \times 10^{-5}$; positive regulators: ANOVA, $N = 2{,}759$, df = 5, $F = 2.75$ $P = 0.01$; negative regulators $N = 2{,}759$, df = 5, $F = 7.98$ $P = 10^{-7}$. Similar results hold when nontransferred genes are excluded from the analysis (data not shown). Age of transfer is given as the number of branching points when traversing down the phylogenetic tree from *Escherichia coli* K12 (fig. 1).

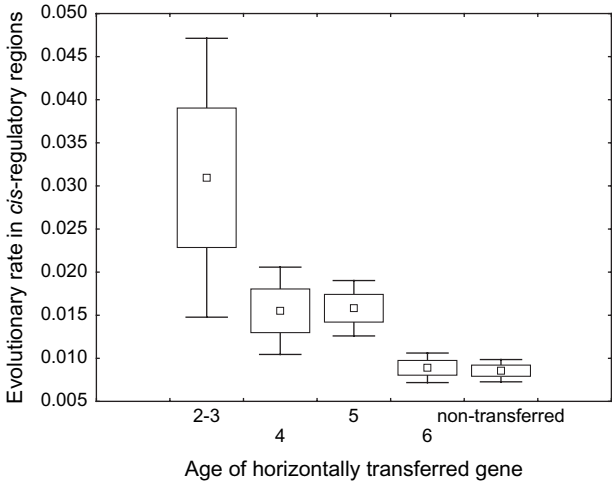

FIG. 4.—Rate of evolution at upstream noncoding DNA regions as a function of age of transfer. Central squares are mean, upper and lower box margins indicate SE, whiskers indicate 2 × SE. ANOVA, $N = 875$, df = 4, $F = 14.55$, $P = 10^{-10}$.

transferred into *E. coli* K12 after its split from the *S. flexneri* lineage were excluded from the analysis. As predicted, the rate of evolution at the 500 noncoding base pairs upstream of transcriptional units—where almost all bacterial transcription control elements are found—gradually decreases with the age of horizontal transfer (fig. 4; this result is not affected by details of the alignment of regulatory regions or the substitution rate calculations, supplementary table 4 [Supplementary Material online]).

Coexpression Evolution of Interacting Proteins after Horizontal Transfer

Improper regulation will be especially problematic for genes involved in protein–protein interactions, as low dosage impedes the formation of proper complex formation (Deutschbauer et al. 2005), and imbalance of complex subunits can lead to harmful protein aggregation (Papp et al. 2003). Thus, even when suitable interaction partners are present in the host of a newly transferred gene, protein expression of the partners needs to be synchronized to enable efficient cofunction (Jain et al. 1999).

To investigate the coevolution of regulatory and protein–protein interaction networks, we analyzed 7,605 interacting gene pairs with at least one member that has not undergone HGT into the *E. coli* lineage over the evolutionary time scale analyzed here. For all gene pairs, we calculated Pearson's correlation coefficient (r) of the mRNA expression vectors across 52 different biological conditions.

In agreement with expectations, coexpression is strikingly poor for protein pairs where one partner was recently transferred into *E. coli* K12. Synchronized regulation improves over evolutionary time, until it reaches levels comparable with those of nontransferred gene pairs (fig. 5).

The trends of regulatory change over evolutionary time observed in figures 3–5 are unlikely to be explained by the gradual loss of nonfunctional horizontally derived genes, for 2 reasons. First, systematic variation extends beyond the *E. coli–Salmonella* split (age classes 5–6), whereas nonfunctional genes are unlikely to persist for such periods in free-living bacterial genomes. Second, excluding computationally predicted pseudogenes has no effect on our results (supplementary table 5, Supplementary Material online).

Conclusions

Recently transferred genes have few physical and regulatory interactions. This indicates that high connectivity acts as a barrier to horizontal gene flow not only in protein interaction networks (see also Wellner et al. 2007) but also in regulatory networks. We demonstrate here that the evolutionary fine-tuning of transcriptional regulation is exceedingly slow. The integration into the existing regulatory circuits of *E. coli* is still ongoing for many genes that entered the ancestral genome long before the diversification of the *E. coli* lineages 8–22 MYA (Bergthorsson and Ochman 1998).

The barrier to gene flow associated with complex regulatory interactions appears "soft": We observe a slow but steady integration of transferred genes into the regulatory system of their host, reflected in accelerated evolution of regulatory sites, increasing complexity of *cis*-regulatory interactions, and improving coregulation of physically interacting proteins. Our results thus reinforce the view that during the evolution of bacterial regulatory circuits, transcription factors and their binding targets can evolve largely independently, allowing genes to join or leave regulons depending on environmental circumstances (Madan Babu et al. 2006).

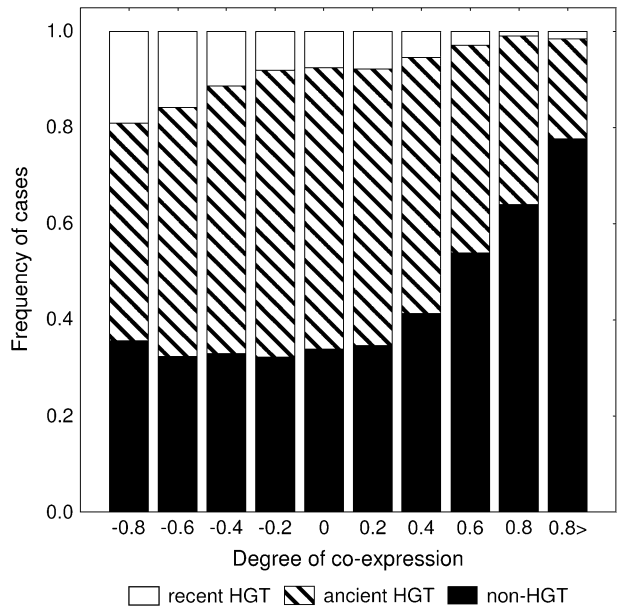

FIG. 5.—Coexpression of interacting protein pairs is poor immediately after HGT but improves over evolutionary time. We concentrated on interacting gene pairs with at least one member that has not undergone HGT into the *Escherichia coli* lineage over the evolutionary time scale analyzed here. The data set was divided into 3 categories based on the transfer status of the other member (nontransferred, product of recent [branches 1–3], or ancient [4–6] transfer events). For all gene pairs, we calculated Pearson's correlation coefficient (r) of the mRNA expression vectors across 52 different biological conditions. The figure shows relative proportions of all 3 categories. We observe a gradual increase in the frequency of nontransferred genes ($N = 7{,}605$, $\chi^2 = 463$, df = 9, $P < 10^{-93}$). The fraction of recently transferred genes among all transferred genes decreases drastically with increasing coexpression ($N = 4{,}431$, $\chi^2 = 91$, df = 9, $P < 10^{-16}$).

A recent publication (Lagomarsino et al. 2007) concludes that duplication of genes encoding transcription factors has played an important role in the evolution of the *E. coli* transcription network, whereas horizontally transferred genes were mostly added at the bottom layer of the network. Given that higher levels of the regulatory hierarchy may not only require many regulatory interactions but may also influence a wide range of downstream functions, this should not be surprising: Perturbation of higher regulatory levels by horizontally transferred genes is less likely to be selectively favorable compared with downstream targets. In qualitative agreement with this idea, we find that those transcription factors transferred into the *E. coli* lineage have on average less than half as many targets compared with nontransferred transcription factors (nontransferred: 14.3 ± 4.9 targets [$N = 35$]; transferred: 6.1 ± 0.7 [$N = 85$]; $P = 4.5 \times 10^{-6}$ from Mann–Whitney U test).

In contrast to the evolution of the regulatory network, we find that even proteins resulting from ancient transfers have few direct interactions with other proteins. This indicates that physical interactions provide a "hard" barrier to the fixation of horizontally transferred genes: Interaction partners have to be present immediately in the new host for the transferred gene to provide a selective advantage. As hubs of the protein–protein interaction network are likely to be relatively unspecific in their binding, our results are consistent with the finding that newly acquired genes in *E. coli* frequently attach to such hubs (Ochman et al. 2007). Based on indirect evidence from the comparison of 2 complementary methods of HGT identification, Wellner et al. (2007) have recently suggested that protein–protein interactions are slowly built over evolutionary time, analogous to our findings on regulatory network evolution. Although we see some weak trends in this direction, our direct evidence does not lend statistically significant support to that idea.

One obvious way to circumvent the barrier to horizontal gene flow imposed by protein interactions is the cotransfer of interacting partners. Frequently, transferred stretches of DNA indeed contain complete operons encoding large supramolecular structures rather than just individual genes (Lawrence 1997; Homma et al. 2007). As interacting proteins often reside in the same operon (or are encoded by neighboring genes) (Dandekar et al. 1998; Homma et al. 2007), the loss of interaction partners might hence be evaded: transfer of complete operons conserves both physical partnership and coregulation. There are several clear examples of cotransfers of operon parts encoding protein complex subunits (supplementary table 6, Supplementary Material online). In a similar vein, transcription factors in *E. coli* (but not in eukaryotic yeast) frequently regulate target genes that sit adjacent to them in the genome (Hershberg et al. 2005), thereby facilitating cotransfer of transcription factors and their targets. However, successful transfer and initial integration of foreign DNA become less likely for larger fragments; hence, large modules of interacting genes are unlikely to be transferred in a single step.

Future studies should further characterize how modular organization of different cellular networks might reduce pleiotropic constraints and hence facilitates adaptation to new environmental conditions by horizontal transfer of moderately sized genetic modules (McAdams et al. 2004; Kashtan and Alon 2005; Alm et al. 2006; Homma et al. 2007). Furthermore, it is also unclear how far the low gene dosages of horizontally transferred genes interfere with detecting the number of interacting partners in these genes.

How can the exceedingly slow fine-tuning of regulatory interactions be explained? We speculate that this process involves adjustments at many biochemical levels. For instance, it is well established that usage of rare codons (Kane 1995) and low GC content (Navarre et al. 2006) strongly influence levels of gene expression. Previous genomic studies have demonstrated that evolutionary modification of these parameters is very slow (Lawrence and Ochman 1997), possibly because it requires many nucleotide changes with relatively small individual effects on fitness. We found a lower rate of acquisition of binding sites for negative compared with positive regulators. This finding would be consistent with lower selective pressures on the reduction compared with the enhancement of transcription levels: maybe too much of a protein is often less costly to the organism than too little. This decoupled evolution of up- and downregulation might also contribute to the prolonged regulatory evolution.

It is not clear if the slow rate of regulatory evolution observed here for horizontally transferred genes also

applies to the fine-tuning of regulatory interactions among native genes in response to environmental changes. Environmental changes often happen on shorter time scales than those examined here. Accordingly, such a generalization of our findings would suggest that many interactions in existing regulatory networks might still be evolving toward their optimum.

Supplementary Material

Supplementary tables 1–6 and figures 1–5 are available at *Molecular Biology Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to thank William Martin, Balazs Papp, Laurence Hurst, and Peer Bork for helpful discussions. C.P. is supported by the Hungarian Scientific Research Fund (Hungarian Research Grant). M.J.L. acknowledges financial support from the Deutsche Forschungsgemeinschaft.

Literature Cited

Alm E, Huang K, Arkin A. 2006. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol. 2:e143.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arifuzzaman M, Maeda M, Itoh A, et al. (23 co-authors). 2006. Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. Genome Res. 16:686–691.

Ball CA, Awad IA, Demeter J, et al. (13 co-authors). 2005. The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res. 33: D580–D582.

Bergthorsson U, Ochman H. 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol Biol Evol. 15:6–16.

Bhardwaj N, Lu H. 2005. Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics. 21:2730–2738.

Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci USA. 101:9722–9727.

Brochier C, Philippe H, Moreira D. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet. 16: 529–533.

Butland G, Peregrin-Alvarez JM, Li J, et al. (14 co-authors). 2005. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature. 433: 531–537.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 23:324–328.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. Science. 301:829–832.

Daubin V, Ochman H. 2004. Quartet mapping and the extent of lateral transfer in bacterial genomes. Mol Biol Evol. 21:86–89.

Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics. 169:1915–1925.

Dorman CJ. 2007. H-NS, the genome sentinel. Nat Rev Microbiol. 5:157–161.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Garcia-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10:1719–1725.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19: 2226–2238.

Gralla JD, Collado-Vides J. 1996. Organization and function of transcription regulatory elements in *Escherichia coli* and *Salmonella typhimurium*. In: Neidhardt FC, Curtiss R III, Ingraham J, Lin ECC, Low KB, Magasanik B, Reznikoff W, Schaechter M, Umbarger HE, Riley M, editors. Cellular and molecular biology: *Escherichia coli* and *Salmonella typhimurium*. Washington (DC): American Society for Microbiology. p. 1232–1245.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16: 636–643.

Hershberg R, Yeger-Lotem E, Margalit H. 2005. Chromosomal organization is shaped by the transcription regulatory network. Trends Genet. 21:138–142.

Homma K, Fukuchi S, Nakamura Y, Gojobori T, Nishikawa K. 2007. Gene cluster analysis method identifies horizontally transferred genes with high reliability and indicates that they provide the main mechanism of operon gain in 8 species of gamma-Proteobacteria. Mol Biol Evol. 24:805–813.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA. 96:3801–3806.

Jain R, Rivera MC, Moore JE, Lake JA. 2002. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol. 61:489–495.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol. 6:494–500.

Kashtan N, Alon U. 2005. Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci USA. 102:13773–13778.

Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. 2005. EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res. 33:D334–D337.

Koonin EV. 2003. Horizontal gene transfer: the path to maturity. Mol Microbiol. 50:725–727.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol. 55:709–742.

Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Res. 13: 1589–1594.
Lagomarsino MC, Jona P, Bassetti B, Isambert H. 2007. Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. Proc Natl Acad Sci USA. 104:5516–5520.
Lawrence JG. 1997. Selfish operons and speciation by gene transfer. Trends Microbiol. 5:355–359.
Lawrence JG, Hartl DL, Ochman H. 1991. Molecular considerations in the evolution of bacterial genes. J Mol Evol. 33:241–250.
Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 44:383–397.
Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA. 95: 9413–9417.
Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.
Madan Babu M, Teichmann SA, Aravind L. 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol. 358:614–633.
McAdams HH, Srinivasan B, Arkin AP. 2004. The evolution of genetic regulatory systems in bacteria. Nat Rev Genet. 5:169–178.
Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. 3:2.
Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 36:760–766.
Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. 313:236–238.
Nei M, Kumar S. 2000. Molecular evolution andf phylogenetics. Oxford: Oxford University Press.
Nesbo CL, Boucher Y, Doolittle WF. 2001. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol. 53:340–350.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405:299–304.
Ochman H, Liu R, Rocha EP. 2007. Erosion of interaction networks in reduced and degraded genomes. J Exp Zoolog B Mol Dev Evol. 308:97–103.
Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 37:1372–1375.
Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature. 424:194–197.
Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA. 95:6239–6244.
Salgado H, Gama-Castro S, Peralta-Gil M, et al. (12 co-authors). 2006. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. 34:D394–D397.
Siddharthan R. 2006. Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. BMC Bioinformatics. 7:143.
Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12:17–25.
Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, Takahashi N, Isobe T. 2004. Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. Mol Cell Proteomics. 3:780–787.
Wang J, Keightley PD, Johnson T. 2006. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. BMC Bioinformatics. 7:292.
Wellner A, Lurie MN, Gophna U. 2007. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. 8:R156.
Zhaxybayeva O, Nesbo CL, Doolittle WF. 2007. Systematic overestimation of gene gain through false diagnosis of gene absence. Genome Biol. 8:402.

Takashi Gojobori, Associate Editor

Accepted December 18, 2007