Molecular Biology and Evolution

The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer

Ofir Cohen,[1] Uri Gophna,[2] and Tal Pupko*[,1,3]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[2]Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[3]National Evolutionary Synthesis Center, Durham, NC

***Corresponding author:** E-mail: talp@post.tau.ac.il.

Associate editor: Andrew Roger

Abstract

Horizontal gene transfer (HGT) is a prevalent and a highly important phenomenon in microbial species evolution. One of the important challenges in HGT research is to better understand the factors that determine the tendency of genes to be successfully transferred and retained in evolution (i.e., transferability). It was previously observed that transferability of genes depends on the cellular process in which they are involved where genes involved in transcription or translation are less likely to be transferred than metabolic genes. It was further shown that gene connectivity in the protein–protein interaction network affects HGT. These two factors were shown to be correlated, and their influence on HGT is collectively termed the "Complexity Hypothesis". In this study, we used a stochastic mapping method utilizing advanced likelihood-based evolutionary models to quantify gene family acquisition events by HGT. We applied our methodology to an extensive across-species genome-wide dataset that enabled us to estimate the overall extent of transfer events in evolution and to study the trends and barriers to gene transferability. Focusing on the biological function and the connectivity of genes, we obtained novel insights regarding the "complexity hypothesis." Specifically, we aimed to disentangle the relationships between protein connectivity, cellular function, and transferability and to quantify the relative contribution of each of these factors in determining transferability. We show that the biological function of a gene family is an insignificant factor in the determination of transferability when proteins with similar levels of connectivity are compared. In contrast, we found that connectivity is an important and a statistically significant factor in determining transferability when proteins with a similar function are compared.

Key words: complexity hypothesis, protein interaction network, informational and operational genes, phyletic pattern, genome evolution, horizontal gene transfer.

Research article

Introduction

Comparative genomics have revealed vast and surprising variability in gene content even among closely related species (Berg and Kurland 2002; Mira et al. 2002; Konstantinidis and Tiedje 2004; Koonin and Wolf 2008). The dynamics of genomes remodeling include drastic genome erosions by gene losses (Moran et al. 2009) and acquisition of novel genetic material by gene gains through horizontal gene transfer (HGT) (Syvanen 1994; Hacker and Carniel 2001). A pivotal role for HGT was demonstrated in the adaptation of organisms to new ecological niches (Gogarten and Townsend 2005), acquisition of novel functions (Pennisi 2004; Gogarten and Townsend 2005), metabolic networks expansions (Pal et al. 2005), and speciation (Lawrence 1999). The transfer of genes among bacteria also bears significant medical implications as the emergence of new virulent strains as well as their resistance to antibiotics is mainly attributed to HGT (Holden et al. 2004; Gal-Mor and Finlay 2006). Thus, studying HGT dynamics and the factors that determine gene transferability is important for evolutionary, ecological, and molecular biology studies.

Although most genes are susceptible to HGT (Sorek et al. 2007), it is well established that the tendency to undergo HGT is highly variable among genes (Nakamura et al. 2004; Cohen et al. 2008; Hao and Golding 2008). Over a decade ago, it was suggested that the biological process in which a gene is involved strongly affects its transferability. It was shown that informational genes are less transferable than operational genes. Later, it was additionally shown that the number of protein–protein interactions (PPIs) is an important factor in determining transferability. The dependency of transferability on these two factors, the biological process and the network connectivity, is now collectively referred to as the "complexity hypothesis" (Rivera et al. 1998; Doolittle 1999; Jain et al. 1999, 2002; Sicheritz-Ponten and Andersson 2001; Gogarten et al. 2002; Brown 2003; Wellner et al. 2007; Lercher and Pal 2008).

Since it was suggested, the complexity hypothesis was in the center of the discussion regarding gene transferability: The hypothesis was extended (Aris-Brosou 2005) and

received support from both bioinformatic analyses (Lercher and Pal 2008) and experimental studies (Wellner and Gophna 2008). Nevertheless, it was also debated and criticized (Brochier et al. 2000; Nesbo et al. 2001).

Testing the validity of the complexity hypothesis requires accurate inference of HGT events. There are three widely used computational approaches to detect HGT, each tailored toward the detection of only a subset of all transfer events. The first detects genes with phylogenetic incongruence as compared with the inferred ribosomal trees or trees that are supposed to represent the organismal evolutionary history. This approach is only suitable for relatively widespread genes with "not too much or too little" sequence divergence (e.g., Graybeal 1994). The second detects genes that are significantly different from the rest of the genome in some compositional attributes such as G+C content or codon usage. This approach can only detect recent transfer events due to sequence amelioration (Koski et al. 2001; Wang 2001). The third uses a presence-absence matrix of gene families across multiple genomes (phyletic pattern) to detect acquisition events of gene families along the assumed phylogeny. Although this approach is suitable for the detection of both recent and ancient events of all gene families, it is only capable of detecting transfer events that resulted in the acquisition of the first copy of a particular gene family. For example, this approach ignores xenologous gene replacements or HGT events that result in additional paralogs. This subset of transfer events may be only a fraction of all HGTs, but it is of a particular evolutionary significance as the acquisition of a novel gene family increases the proteomic repertoire of the recipient and holds the greatest potential for functional innovations and adaptations.

HGT inference from phyletic patterns has been classically inferred based on the parsimony criterion (Yang 1996; Mirkin et al. 2003; Cordero et al. 2008; Lercher and Pal 2008). Recently, more statistically robust models for phyletic pattern analysis were developed in which the dynamics of gain and loss of gene families is modeled as a stochastic process (Hao and Golding 2006). Several improvements to such evolutionary models were developed (Cohen et al. 2008; Hao and Golding 2008; Spencer and Sangaralingam 2009; Cohen and Pupko 2010), and we have utilized these maximum-likelihood models to develop a methodology to accurately detect branch-specific HGT events (Cohen and Pupko 2010).

Here, we apply our HGT detection methodology to characterize the factors that determine transferability in genome-wide data. Specifically, we test the complexity hypothesis and disentangle cellular function and the number of protein interactions as factors that determine transferability. Notably, in this manuscript, for brevity, we use the general term HGT in lieu of the more accurate expression, gene family acquisition by HGT. Thus, our conclusions regarding the complexity hypothesis are limited to this type of HGT events.

Methods

Phyletic Pattern and Phylogeny

The presence-absence matrix of gene family was extracted from the Clusters of Orthologous Groups of proteins (COG) database (Tatusov et al. 2003), which contains 4,873 gene families across 66 species (50 bacteria, 13 archaea, and 3 fungi). In this research, we focused on HGT inference from the set of 50 bacterial genomes. Previous research has shown that gain and loss dynamics is different in parasitic bacteria versus free living organisms (Spencer and Sangaralingam 2009). Because the models used here do not allow branch-specific changes in the evolutionary process, we removed the 12 known parasitic bacteria from our analysis (*Mycoplasma pulmonis*, *Mycop. pneumoniae*, *Mycop. genitalium*, *Ureaplasma urealyticum*, *Buchnera* sp. *APS*, *Rickettsia prowazekii*, *R. conorii*, *Treponema pallidum*, *Borrelia burgdorferi*, *Chlamydia trachomatis*, *Chlamydophila pneumoniae*, and *Mycobacterium leprae*), retaining 38 bacterial genomes. The COG data set definition of gene family requires its presence in at least three genomes. Therefore, the exclusion of species from the original COG database dictates retaining only gene families that are present in at least three genomes within our data set. After this filtering criterion, 3,915 gene families were retained.

The analysis was based on the assumed tree topology of Ciccarelli et al. (2006), which was constructed from a set of "core" genes that are assumed to be resistant to gene transfer. As a control, the analysis was repeated with a topology constructed based on ribosomal RNA (rRNA) sequences (Yarza et al. 2008). In both cases, branch lengths are re-estimated from the phyletic pattern using the evolutionary models.

Inference of HGT Events Using Stochastic Mapping

The gain and loss dynamics were modeled using gain loss mixture model (Cohen and Pupko 2010) in which variability in the gain and loss rates is allowed among gene families. Based on the evolutionary model and the assumed phylogeny, the stochastic mapping approach (Minin and Suchard 2008) allows for the inference of gain and loss events for each gene family along each branch (Cohen and Pupko 2010). This methodology allows for the computation of both the expected number of events and the probability of occurrence of gain and loss events.

The overall tendency of a gene family to undergo acquisition by HGT is measured by the posterior expectation of the number of gain events over all branches. We classify gene families to either transferable or not. Transferable gene families are those for which there is a high probability of HGT events during their evolution. To be conservative, we demand a gain event (HGT) in at least two branches as described in a previous research (Cohen and Pupko 2010). The transferability cutoff value is determined by limiting the number of false-positive predictions of gain events to 5% based on simulations. In this study, the transferability cutoff corresponds to a posterior probability of 0.25 for

a gain event. Notably, the cutoff values that result with 5% false-positive predictions may vary with respect to simulation assumptions (Cohen and Pupko 2010). Thus, in this study, a relative strict cutoff value was used, which may result with less than 5% of false positives under realistic assumptions. Moreover, the computations were repeated with both more strict and more permissive cutoff values.

We classify all gain events to either ancient or recent. Recent gains are those that are mapped to external branches, (i.e., branches leading to extant organisms). Gain events mapped to all other branches are considered ancient (i.e., gain events mapped to internal branches).

Network and Protein Interactions

The PPI network and the number of interactions for each gene family were extracted from the STRING database version 8.3 (Jensen et al. 2009). This comprehensive PPI network is based on known interactions from several databases covering several model organisms (Salwinski et al. 2004; Alfarano et al. 2005; Joshi-Tope et al. 2005; Chatr-aryamontri et al. 2007; Kerrien et al. 2007; Vastrik et al. 2007; Breitkreutz et al. 2008; Kanehisa et al. 2008) and is augmented by methods that accurately predict interactions (Harrington et al. 2008; Skrabanek et al. 2008). As a control, the PPI network of the Database of Interacting Proteins (DIP) (Xenarios et al. 2000; Salwinski et al. 2004) version June 2010 was used. Unlike STRING, this network only considers experimentally validated interactions (i.e., it does not include predicted interactions).

Interactions in the STRING data set are given confidence score based on benchmarking with manually curated interaction maps (Kanehisa et al. 2008) in which for each pair of gene families the interaction confidence is denoted by a value in the range 0–1,000. Protein families in which all interactions have a zero confidence score may reflect lack of data rather than genuine non-interacting protein families. These were excluded from the analysis resulting in 2,442 gene families. Notably, in our analyses, we only consider interactions with a confidence score above a certain threshold. For example, a protein family having one reported interaction with a confidence score of 400 will be analyzed as having a single interaction when the threshold is 150 and zero interactions when the threshold is 700.

Functional Categories

The biological process in which each gene family is involved (functional category) is extracted from the COG database in which there are 25 specific categories grouped into four meta-categories. We limited our analysis to functional categories with at least three gene families. Thus, we retain the four meta-categories (Information storage and processing, Cellular processes and signaling, Metabolism, and Poorly characterized) and 20 specific categories (Translation, ribosomal structure, and biogenesis; Transcription; Replication, recombination, and repair; Cell cycle control, cell division, and chromosome partitioning; Defense mechanisms; Signal transduction mechanisms; Cell wall/membrane/envelope biogenesis; Cell motility; Intracellular trafficking, Secretion, and vesicular transport; Posttranslational modification, protein turnover, and chaperones; Energy production and conversion; Carbohydrate transport and metabolism; Amino acid transport and metabolism; Nucleotide transport and metabolism; Coenzyme transport and metabolism; Lipid transport and metabolism; Inorganic ion transport and metabolism; Secondary metabolites biosynthesis, transport, and catabolism; General function prediction only; and Function unknown). In the analysis of functional categories, 158 gene families have more than one functional category label. These gene families were included independently in each functional category analysis.

Statistical Analysis of Function and Transferability Association

To test for association between a functional category and transferability, we computed the ratio between the fraction of transferable genes in this category to the fraction of transferable genes not in this category. We term this ratio "relative transferability," which is equivalent of the often used term "relative risk." A relative transferability significantly higher than one suggests a higher propensity for a gene family to be transferred when included in this functional category compared with all other functional categories. Statistical significance is determined using Fisher's exact test. The classification of a gene family as transferable is based on stochastic mapping (see above).

We additionally compute the relative transferability while accounting for variable levels of connectivity. Specifically, we treated the data as stratified by the number of PPIs. We thus computed the relative transferability in each functional category accounting for this stratification using the Mantel–Haenszel test. Gene families were stratified into 45 levels of connectivity, in which each stratum has at least ten gene families. Notably, similar results were obtained when that data were stratified to 94 levels of connectivity (at least three gene families in each stratum) or to seven levels of connectivity (at least 100 gene families in each stratum). This indicates that the results are highly robust to the choice of stratification resolution (data not shown).

The stratification of the gene family according to their connectivity was done as follows. All gene families are sorted according to their connectivity (number of PPIs). The first stratum comprises the group of gene families with the lowest number of interactions. We incrementally add gene families with the next lower levels of connectivity to this stratum until at least ten gene families are included. All gene families with the exactly the same level of connectivity are added to a stratum, even if the size increases above 10. Once a stratum is defined, we build the next stratum.

Results and Discussion

High Number of Protein Interactions Acts As Barrier to HGT

Gain and loss dynamics of gene families were studied using the gain loss mixture model (Cohen and Pupko 2010). The ML estimate for each of the model parameters is given in

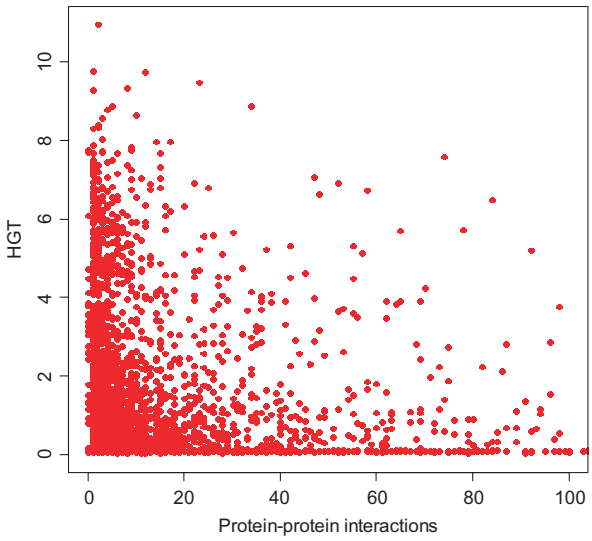

FIG. 1. HGT as a function of PPIs. Each dot represents one gene family. The *X* axis is the overall number of PPIs of the gene family with other gene families. The *Y* axis corresponds to the posterior number of HGT events.

supplementary table S1, Supplementary Material online. This model was used to infer gain (HGT) events for each gene family and for each branch using stochastic mapping. It was previously shown that the number of protein interactions (connectivity) is associated with gene family transferability by comparing transferability of genes with low versus high connectivity levels (Wellner et al. 2007; Davids and Zhang 2008; Lercher and Pal 2008). To gain further insight regarding this "connectivity barrier" (i.e., protein interactions that hinder or reduce transferability), instead of using such a binning approach, we directly computed the correlation between connectivity and transferability (fig. 1). Our analysis indicates that more HGT events are expected for protein families with low connectivity levels (Spearman coefficient computed for all gene families $R = -0.422$, P value $< 8.18 \times 10^{-105}$, table 1).

Each interaction is given a confidence value (see Methods). We performed the same computation with different cutoff levels for the inclusion of interactions. In all cases, the negative correlation was highly significant (supplementary table S2A, Supplementary Material online). Notably, increasing the threshold reduces the number of included interactions (overall number of network interactions 62,642 and 5,861 for the lowest and highest confidence threshold, respectively) and the corresponding R coefficients ($R = -0.422$ and -0.298 for the lowest and the highest confidence threshold, respectively). Although the increase in the confidence cutoff may have reduced the number of false interactions, the decrease in the correlation strength suggests that using a stringent cutoff value significantly reduced the number of true interactions as well. Notably, we found very similar correlation levels between HGT and connectivity with exclusively low and mid confidence levels. These results (supplementary table S2B, Supplementary Material online) suggest that with the STRING confidence score method, even interactions with low confidence contribute to the connectivity barrier.

It may be claimed that the connectivity barrier is mainly the result of the most extreme cases in the connectivity spectrum, that is, that the majority of the signal arises from isolated gene families and hub gene families. We repeated the correlation analysis, removing from the analysis gene

Table 1. The Spearman's Correlation (R) between Connectivity and Transferability Computed Separately for Various Functional Categories.

Functional Category	R	P	Number of Gene Families
All	−0.422	8.18×10^{-105}	2,442
Information storage and processing	−0.518	1.08×10^{-26}	382
Cellular processes and signaling	−0.353	5.95×10^{-15}	487
Metabolism	−0.39	2.94×10^{-37}	1,017
Poorly characterized	−0.244	2.35×10^{-09}	626
Translation, ribosomal structure, and biogenesis	−0.213	0.0112	150
Transcription	−0.349	4.78×10^{-04}	105
Replication, recombination, and repair	−0.444	1.98×10^{-07}	133
Cell cycle control, cell division, and chromosome partitioning	−0.464	0.00727	35
Defense mechanisms	−0.474	0.00723	34
Signal transduction mechanisms	−0.235	0.0265	93
Cell wall/membrane/envelope biogenesis	−0.289	9.72×10^{-04}	138
Cell motility	−0.255	0.0577	57
Intracellular trafficking, secretion, and vesicular transport	−0.331	0.0105	63
Posttranslational modification, protein turnover, and chaperones	−0.411	1.03×10^{-05}	115
Energy production and conversion	−0.407	2.50×10^{-08}	185
Carbohydrate transport and metabolism	−0.461	2.35×10^{-09}	162
Amino acid transport and metabolism	−0.387	6.65×10^{-09}	223
Nucleotide transport and metabolism	−0.471	1.79×10^{-05}	81
Coenzyme transport and metabolism	−0.513	5.47×10^{-10}	139
Lipid transport and metabolism	−0.207	0.0831	71
Inorganic ion transport and metabolism	−0.174	0.0359	151
Secondary metabolites biosynthesis, transport, and catabolism	−0.329	0.0203	52
General function prediction only	−0.299	1.57×10^{-07}	314
Function unknown	−0.152	0.00968	312

NOTE.—The data were partitioned into groups of gene families based on the COG functions. The P values were corrected for multiple testing using false discovery rate method (Benjamini and Hochberg 1995).

Table 2. Connectivity and HGT Propensity of All Gene Families and Specific Functional Categories.

Functional Category	Mean HGT	SE HGT	Mean PPI	SE PPI
All	1.731	0.04	25.65	1.02
Information storage and processing	1.188	0.1	56.5	3.89
Cellular processes and signaling	1.54	0.08	26.45	2.94
Metabolism	1.781	0.06	21.16	0.97
Poorly characterized	2.107	0.08	16.75	2.2
Translation, ribosomal structure, and biogenesis	0.469	0.08	84.51	6.22
Transcription	1.374	0.17	43.35	9.45
Replication, recombination, and repair	1.856	0.21	47.03	7.56
Cell cycle control, cell division, and chromosome partitioning	1.704	0.35	16.8	3.62
Defense mechanisms	2.49	0.47	15.94	5.35
Signal transduction mechanisms	1.559	0.16	33.35	9.14
Cell wall/membrane/envelope biogenesis	1.671	0.16	16.97	2.49
Cell motility	1.008	0.16	15.12	2.62
Intracellular trafficking, secretion, and vesicular transport	0.961	0.14	14.19	2.54
Posttranslational modification, protein turnover, and chaperones	1.394	0.16	45.41	8.98
Energy production and conversion	1.952	0.14	24.41	2.42
Carbohydrate transport and metabolism	2.309	0.17	19.87	2.21
Amino acid transport and metabolism	1.625	0.12	23.91	2.34
Nucleotide transport and metabolism	1.336	0.2	33.69	5.66
Coenzyme transport and metabolism	1.337	0.15	17.39	1.79
Lipid transport and metabolism	1.287	0.18	31.99	4.04
Inorganic ion transport and metabolism	1.889	0.16	15.61	2.37
Secondary metabolites biosynthesis, transport, and catabolism	2.137	0.29	18.08	4.67
General function prediction only	1.998	0.11	26.33	4.28
Function unknown	2.217	0.11	7.112	0.67

NOTE.—The PPI (connectivity) and HGT values are computed for each functional category and for all gene families as reference. SE, standard error.

families with less than one and higher than 50 interactions, respectively. This analysis shows that the connectivity is also informative with respect to HGT for intermediate level of interactions ($R = -0.362$, P value $< 9.2 \times 10^{-66}$).

The Biological Functional As a Factor Determining HGT Extent

Others and we have previously shown that the biological function of a gene family is important in determining its propensity to undergo HGT (Rivera et al. 1998; Nakamura et al. 2004; Merkl 2006; Choi and Kim 2007; Hao and Golding 2008; Kanhere and Vingron 2009; Cohen and Pupko 2010). Here, we show that the mean number of HGT events dramatically changes among various functional categories (table 2, HGT columns). In agreement with previous studies, the lowest HGT levels are observed for the informational genes (involved in transcription and translation), where the most pronounced trend was found in genes associated with the ribosome and related with translation (COG functional category: "translation, ribosomal structure, and biogenesis"). The average expected number of HGT events per gene family in this category was below 0.47, substantially lower than the 1.73 events per gene family, which is the average over all gene families. Statistical significant differences were found among the 20 specific functional categories and also among the four "meta-categories" (P values $< 1.7 \times 10^{-40}$ and 4.52×10^{-34}, Kruskal–Wallis test, respectively).

We further studied the association between functional category and the propensity for HGT by computing the relative transferability factor of each functional category (see Methods for more details). In table 3, we summarize the relative transferability of all functional categories and find that several functional categories have relative transferability that is significantly different than one. In agreement with the lower computed average HGT, the relative transferability value of the function "translation, ribosomal structure, and biogenesis" is 0.276, which is highly significant even after correction for multiple testing (P value $< 3.55 \times 10^{-6}$, Fisher's exact test).

The classification of a gene family as transferable is dependent on an estimated "transferability cutoff" (see Methods). To verify that the obtained results are robust in this respect, we perform additional computations with both more strict and more permissive cutoffs. Changing the cutoff substantially affects the estimation of the overall percentage of transferable genes from 32.31% to 23.91% and 42.51% for the more strict and more permissive cutoffs, respectively. However, the relative transferability factors of the various functional categories were very similar (supplementary tables S3A and B, Supplementary Material online).

Disentangling Biological Functional and Connectivity in Determining HGT Frequency

The complexity hypothesis relates two biological factors to the tendency of gene families to undergo HGT—the connectivity and the function (or biological process). Our results above show that HGT is influenced by each of these factors when analyzed separately. However, the various functional categories vary in terms of their connectivity. In table 2, we show the average connectivity of each functional category. Substantial differences are observed among the functional categories, with averages ranging from as high as 84.51 for the "translation, ribosomal structure,

Table 3. The Relative Transferability of Gene Families in Each Functional Category.

Functional Category	Relative Transferability	*P*	Relative Transferability (MH)	*P* value (MH)
Information storage and processing	0.608	0.00104	0.774	0.344
Cellular processes and signaling	0.866	0.408	0.874	0.617
Metabolism	1.078	0.614	1.138	0.498
Poorly characterized	1.314	0.0216	1.072	0.971
Translation, ribosomal structure, and biogenesis	0.276	2.84×10^{-06}	0.418	0.0864
Transcription	0.668	0.318	0.73	0.617
Replication, recombination, and repair	1.05	0.9	1.241	0.617
Cell cycle control, cell division, and chromosome partitioning	0.883	0.903	0.899	0.977
Defense mechanisms	1.094	0.88	0.955	0.977
Signal transduction mechanisms	0.861	0.782	0.905	0.977
Cell wall/membrane/envelope biogenesis	1.13	0.726	1.116	0.977
Cell motility	0.321	0.0263	0.314	0.0909
Intracellular trafficking, secretion, and vesicular transport	0.633	0.408	0.564	0.349
Posttranslational modification, protein turnover, and chaperones	0.8	0.596	0.899	0.977
Energy production and conversion	1.206	0.408	1.33	0.344
Carbohydrate transport and metabolism	1.37	0.156	1.364	0.344
Amino acid transport and metabolism	0.984	0.943	1.067	0.977
Nucleotide transport and metabolism	0.836	0.782	0.962	0.977
Coenzyme transport and metabolism	0.769	0.408	0.79	0.617
Lipid transport and metabolism	0.78	0.614	0.928	0.977
Inorganic ion transport and metabolism	1.027	0.903	0.98	0.977
Secondary metabolites biosynthesis, transport, and catabolism	1.134	0.853	1.073	0.977
General function prediction only	1.18	0.408	1.059	0.977
Function unknown	1.334	0.0587	1.057	0.977

NOTE.—Relative transferability refers to the fraction of transferable gene families within each functional category divided by the fraction of transferable gene families among other gene families. This computation is repeated, once when all connectivity levels are aggregated and once when accounting for connectivity stratification using Mantel–Haenszel (MH) test. The *P* values were corrected for multiple testing using the false discovery rate method (Benjamini and Hochberg 1995).

and biogenesis" category to only 7.11 for the "function unknown" category. This observed difference among categories is statistically significant in the comparison of both the specific categories and the meta-categories (*P* values $<7.77 \times 10^{-60}$ and 5.48×10^{-60}, Kruskal–Wallis test, respectively). Given this strong association between functional category and connectivity, HGT dependence on the function may be a side effect of this variance in connectivity. Alternatively, it is possible that the observed effect of connectivity over HGT propensity is a by-product of the differences in functionality. Here, we tried to test for the effect of each of these factors, controlling for the effect of the other.

We computed the correlation between connectivity and transferability for each functional group separately. Our results show that the connectivity barrier holds even when the functional category factor is accounted for. Specifically, for the vast majority of functional categories, a significant negative correlation was observed between connectivity and transferability (table 1). However, the impact of the connectivity barrier was different across functional groups. We found that connectivity was the most influential in informational genes with Spearman's coefficient of −0.518, while in both metabolic and cellular groups, the coefficients were lower: −0.39 and −0.353, respectively. The lowest correlation between connectivity and transferability was found for the "poorly characterized" meta-category and for the "function unknown" category with Spearman's coefficients of −0.244 (*P* value 2.35×10^{-09}) and −0.152 (*P* value 0.0096), respectively. The only two functional categories in which the correlation was not significant after correction for multiple testing are "cell motility" and "lipid transport and metabolism".

We next tested if the biological function is a determining factor for transferability when controlling for the connectivity level. We thus computed the relative transferability in each functional category accounting for different levels of connectivity using Mantel–Haenszel test (see Methods). Our results show that when controlling for connectivity, the impact of functional category on transferability drastically diminishes and becomes not significant for all the functional categories (table 3). For example, when accounting for connectivity levels, the relative transferability of informational genes is raised from 0.61 to 0.82. Similarly, for poorly characterized genes, the relative transferability decreases from 1.31 to 1.03. Importantly, using Mantel–Haenszel test, after correction for multiple testing, none of the functional categories is found to have relative transferability that is significantly different from one. The only exception is the functional category "translation, ribosomal structure, and biogenesis" in which the relative transferability is significantly lower than one when the more permissive criterion for transferability is used (supplementary table S3B, Supplementary Material online). This result is not surprising because these gene families are known to be among the so called "core" of the genome, which is highly resistant to HGT (e.g., Ciccarelli et al. 2006; Sorek et al. 2007). To conclude, these results demonstrate that when the connectivity level is taken into account, the functional category is not a significant factor in determining the propensity of gene families to undergo HGT events.

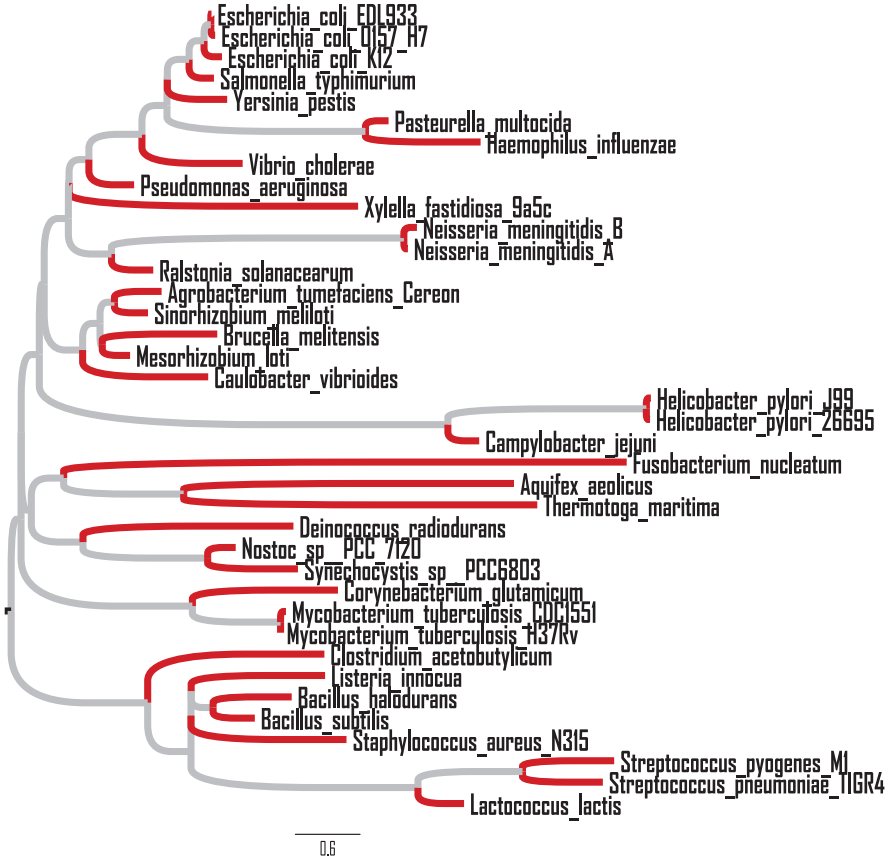

FIG. 2. The phylogeny with branches color-coded as recent or ancient. The phylogenetic tree used in this research. Recent branches are colored red and ancient branches are colored gray.

The Connectivity Barrier Holds Both for Recent and Ancient Acquisitions

The stochastic mapping methodology infers branch-specific gain events. We tested whether the connectivity barrier exists both for recent and for ancient transfers by partitioning the branches of the tree to two groups, recent and ancient. Figure 2 depicts the phylogeny used in this research with branches color-coded as either recent or ancient. Our results show that this is indeed the case: connectivity is a strong predictor for transferability for both recent and ancient HGT events with Spearman's coefficients of -0.39 and -0.43, and P values of 6.01×10^{89} and 2.9×10^{-108}, respectively. Notably, protein interaction data were derived from contemporary experimental observations, and thus, our observations show that current information regarding connectivity is highly informative for ancient HGT events. These results may be explained by a slow evolutionary rate of the PPI network, that is, the connectivity of gene families in current microbes highly resemble that of hypothetical ancestral lineages. This interpretation is in agreement with the findings of Lercher and Pal (2008).

Controls and Additional Tests

The above results were validated with respect to several assumptions. First, we inferred HGT events assuming the tree topology of Ciccarelli et al. (2006). The results obtained were qualitatively the same when all computations were repeated assuming the rRNA tree (Yarza et al. 2008), with detailed results in supplementary tables S4 and S5, Supplementary Material online. The Cicarelli tree was chosen to be the main reference as it obtained higher maximum log-likelihood value compared with the rRNA tree (supplementary table S1, Supplementary Material online). Second, connectivity was inferred based on the STRING database (Jensen et al. 2009). The conclusions were essentially the same when interactions were extracted from the DIP database instead (Salwinski et al. 2004), with detailed results in supplementary table S2, Supplementary Material online.

Conclusions

Since it was suggested, the complexity hypothesis was debated: It was shown that for cases of homologous gene acquisition, the complexity barrier may be low (Wellner et al. 2007; Omer et al. 2010). However, here, we demonstrate that gene family acquisition apparently has very different evolutionary characteristics and involves a substantial complexity barrier that is not restricted to particular protein functions. Our results are based on robust statistical models and methodologies and on a large corpus of phyletic data, which are radically different than those that were available when the complexity hypothesis was first suggested. Using these data and methods, we were able to quantify the extent to which HGT of gene families is determined by the functional category and the number of

protein–protein connections that characterize them. When assessing barriers to HGT and the importance of these factors in determining transferability, we found that high connectivity hinders HGT events. Finally, we demonstrated that the functional category of a gene family is an insignificant factor in determining HGT, once the connectivity level factor is neutralized.

This study focused on the elucidation of factors that determine HGT. We note that an interesting direction for future research is to apply the methodology presented here to quantify and characterize gene family loss dynamics, that is, to elucidate the factors that determine the propensity for gene family loss (dispensability). The importance of gene loss in shaping microbial genomes in evolution was studied and quantified both computationally (Charlebois and Doolittle 2004; Csuros and Miklos 2006; Marri et al. 2006; Borenstein et al. 2007; Wapinski et al. 2007) and experimentally (Moran et al. 2009) and both gene function and network connectivity had been suggested to play an important role (Krylov et al. 2003; Pal et al. 2006; Wolf et al. 2006; Ochman et al. 2007; Yosef et al. 2009). Notably, because gene loss dynamics is known to be much more common in parasitic bacteria, models that account for a covarion-like type of evolution with regard to gain and loss parameters (heterotachy) should be more suitable to analyze gene loss dynamics. An important step forward in this direction is the recent work of Spencer and Sangaralingam (2009), which clearly shows that a covarion-type model of evolution can better capture gene gain and loss dynamics when reductive evolution in some lineages is evident.

Another interesting direction for future research is to build evolutionary models that explicitly consider the association between connectivity and the gain (and loss) rates. Such models are becoming more and more interesting as the volume of microbial genomic data accumulates and the knowledge regarding PPI becomes more accurate.

Supplementary Material

Supplementary tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Daniel Yekutieli for his help with the statistical analysis. We thank Matthew Spencer for reviewing this paper and for providing helpful criticism and suggestion that significantly improved this manuscript. We thank Nimrod Rubinstein for critically reading the manuscript. T.P. is supported by a grant from the Israel Science Foundation (878/09) and by the National Evolutionary Synthesis Center (NESCent), National Science Foundation #EF-0905606. O.C. is a fellow of the Edmond J. Safra program in bioinformatics.

References

Alfarano C, Andrade CE, Anthony K, et al. (75 co-authors). 2005. The Biomolecularbimolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 33:D418–D424.

Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol Biol Evol.* 22:200–209.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodological).* 57:289–300.

Berg OG, Kurland CG. 2002. Evolution of microbial genomes: sequence acquisition and loss. *Mol Biol Evol.* 19:2265–2276.

Borenstein E, Shlomi T, Ruppin E, Sharan R. 2007. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. *Nucleic Acids Res.* 35:e7.

Breitkreutz BJ, Stark C, Reguly T, et al. (12 co-authors). 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* 36:D637–D640.

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.

Brown JR. 2003. Ancient horizontal gene transfer. *Nat Rev Genet.* 4:121–132.

Charlebois RL, Doolittle WF. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 14:2469–2477.

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Res.* 35:D572–D574.

Choi IG, Kim SH. 2007. Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A.* 104:4489–4494.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.

Cohen O, Pupko T. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol Biol Evol.* 27:703–713.

Cohen O, Rubinstein ND, Stern A, Gophna U, Pupko T. 2008. A likelihood framework to analyse phyletic patterns. *Philos Trans R Soc Lond B Biol Sci.* 363:3903–3911.

Cordero OX, Snel B, Hogeweg P. 2008. Coevolution of gene families in prokaryotes. *Genome Res.* 18:462–468.

Csuros M, Miklos I. 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Lect Notes Comput Sci.* 3909:206–220.

Davids W, Zhang Z. 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. *BMC Evol Biol.* 8:23.

Doolittle WF. 1999. Lateral genomics. *Trends Cell Biol.* 9:M5–M8.

Gal-Mor O, Finlay BB. 2006. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol.* 8:1707–1719.

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

Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol.* 3:679–687.

Graybeal A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Syst Biol.* 43:174–193.

Hacker J, Carniel E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* 2:376–381.

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.

Hao W, Golding GB. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics.* 9:235.

Harrington ED, Jensen LJ, Bork P. 2008. Predicting biological networks from genomic data. *FEBS Lett.* 582:1251–1258.

Holden MT, Feil EJ, Lindsay JA, et al. (45 co-authors). 2004. Complete genomes of two clinical Staphylococcus aureus strains: evidence

for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A*. 101:9786–9791.
Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. 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.
Jensen LJ, Kuhn M, Stark M, et al. (12 co-authors). 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*. 37:D412–D416.
Joshi-Tope G, Gillespie M, Vastrik I, et al. (13 co-authors). 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. 33:D428–D432.
Kanehisa M, Araki M, Goto S, et al. (11 co-authors). 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 36:D480–D484.
Kanhere A, Vingron M. 2009. Horizontal Gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evol Biol*. 9:9.
Kerrien S, Alam-Faruque Y, Aranda B, et al. (24 co-authors). 2007. IntAct—open source resource for molecular interaction data. *Nucleic Acids Res*. 35:D561–D565.
Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A*. 101:3160–3165.
Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res*. 36:6688–6719.
Koski LB, Morton RA, Golding GB. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol*. 18:404–412.
Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*. 13:2229–2235.
Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol*. 2:519–523.
Lercher MJ, Pal C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol*. 25:559.
Marri PR, Hao W, Golding GB. 2006. Gene gain and gene loss in streptococcus: is it driven by habitat? *Mol Biol Evol*. 23:2379–2391.
Merkl R. 2006. A comparative categorization of protein function encoded in bacterial or archeal genomic islands. *J Mol Evol*. 62:1–14.
Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol*. 56: 391–412.
Mira A, Klasson L, Andersson SG. 2002. Microbial genome evolution: sources of variability. *Curr Opin Microbiol*. 5:506–512.
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.
Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323:379–382.
Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet*. 36:760–766.
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, Liu R, Rocha EP. 2007. Erosion of interaction networks in reduced and degraded genomes. *J Exp Zool B Mol Dev Evol*. 308:97–103.
Omer S, Kovacs A, Mazor Y, Gophna U. 2010. Integration of a foreign gene into a native complex does not impair fitness in an experimental model of lateral gene transfer. *Mol Biol Evol*. 27(11):2441–2445.
Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*. 37: 1372–1375.
Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. 2006. Chance and necessity in the evolution of minimal metabolic networks. *Nature*. 440:667–670.
Pennisi E. 2004. Microbiology. Researchers trade insights about gene swapping. *Science* 305:334–335.
Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A*. 95:6239–6244.
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*. 32:D449–D451.
Sicheritz-Ponten T, Andersson SG. 2001. A phylogenomic approach to microbial evolution. *Nucleic Acids Res*. 29:545–552.
Skrabanek L, Saini HK, Bader GD, Enright AJ. 2008. Computational prediction of protein-protein interactions. *Mol Biotechnol*. 38:1–17.
Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.
Spencer M, Sangaralingam A. 2009. A phylogenetic mixture model for gene family loss in parasitic bacteria. *Mol Biol Evol*. 26: 1901–1908.
Syvanen M. 1994. Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet*. 28:237–261.
Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 4:41.
Vastrik I, D'Eustachio P, Schmidt E, et al. (13 co-authors). 2007. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*. 8:R39.
Wang B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol*. 53:244–250.
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.
Wellner A, Gophna U. 2008. Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. *Mol Biol Evol*. 25:1835–1840.
Wellner A, Lurie MN, Gophna U. 2007. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biol*. 8:R156.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc Biol Sci*. 273:1507–1515.
Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. 2000. DIP: the database of interacting proteins. *Nucleic Acids Res*. 28:289–291.
Yang Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol*. 42:294–307.
Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glockner FO, Rossello-Mora R. 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol*. 31: 241–250.
Yosef N, Kupiec M, Ruppin E, Sharan R. 2009. A complex-centric view of protein network evolution. *Nucleic Acids Res*. 37:e88.