Molecular Biology, Vol. 37, No. 4, 2003, pp. 571–584. Translated from Molekulyarnaya Biologiya, Vol. 37, No. 4, 2003, pp. 674–687.
Original Russian Text Copyright © 2003 by V'yugin, Gelfand, Lyubetsky.

UDC 577.2.08:681.31:575.852

Identification of Horizontal Gene Transfer from Phylogenetic Gene Trees

V. V. V'yugin[1], M. S. Gelfand[2], and V. A. Lyubetsky[1]

[1] *Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 101447 Russia;*
E-mail:vyugin@iitp.ru
[2] *State Research Center GosNIIGenetika, Moscow, 113545 Russia*

Received October 23, 2002

Abstract—We suggest a new procedure to search for the genes with horizontal transfer events in their evolutionary history. The search is based on analysis of topology difference between the phylogenetic trees of gene (protein) groups and the corresponding phylogenetic species trees. Numeric values are introduced to measure the discrepancy between the trees. This approach was applied to analyze 40 prokaryotic genomes classified into 132 classes of orthologs. This resulted in a list of the candidate genes for which the hypothesis of horizontal transfer in evolution looks true.

Key words: horizontal gene transfer, evolutionary event, statistics of evolutionary event search, phylogenetic species tree, phylogenetic protein tree, tree reconciliation

INTRODUCTION

It is well known that phylogenetic trees derived from different protein families of the same organisms are often incongruent and often differ from the phylogenetic species tree. This could be explained by errors in protein (gene) tree generation, e.g., errors caused by the different evolution rates for the same gene in different phylogenetic lines. The tree incongruence can also result from evolutionary molecular-level events that occur in gene history independently of the species divergence. These events include gene duplication, gene loss, and horizontal gene transfer. The trees showing evolution of proteins and genes, as well as the events of gene loss and duplication, have been discussed by us in more detail elsewhere [1, 2].

We aim to develop methods to obtain information about these evolutionary events from the phylogenetic data. In this work we suggest a new approach to select the candidate genes involved in gene transfer in the course of the evolutionary history basing on the incongruence between gene trees and species trees induced by these genes.

The horizontal transfer between bacterial genomes is rather common. A bacterium can insert a gene into its chromosome directly from the environment, through phage infection, or from another bacterium by the plasmids [3–5]. It is important for medical applications that many plasmids carry antibiotic resistance genes and/or the 'pathogenic islands' of toxin-encoding genes, invasion protein-encoding genes, etc. Some researchers consider horizontal gene transfer one of the main factors in microbial evolution [6–9]. The genome of *Escherichia coli* contains up to 18% horizontally transferred genes [10]. In the genome of *Thermotoga maritima* 25% genes are related with the archaebacterial rather than with the bacterial genes; probably, they were acquired by horizontal transfer [3, 11]. Earlier, similar results were obtained for the genome of another bacterium, *Aquifex aeolicus* [12].

The problem of computer-assisted search for the horizontally transferred genes was considered earlier [13, 14]. However, these works suggested no ways to solve it. In this work we present a method to select the candidate horizontally transferred genes. The respective software produces the lists of the genes that induced significant incongruence between the gene phylogenetic trees and species phylogenetic trees. Obviously, further selection should be done by experts basing on analysis of the functions for the selected candidate genes as well as the degree of similarity of the candidate gene with other genes of the studied organism and of the putative source.

After this work already passed peer reviewing, interesting results were published [19]. This work suggests an algorithm for most efficient reconciliation of the phylogenetic pattern (a set of genomes containing a certain gene) into the species tree. The events of gene loss, duplication, and horizontal transfer are considered elementary operations. The frequency of gene loss and the frequency of horizontal transfer in the course of bacterial evolution are shown to be similar. The main aim of the work [19] is to recover the gene set for the latest common ancestor of the studied genes; tree rearrangements caused by the horizontal

0026-8933/03/3704-0571$25.00 © 2003 MAIK "Nauka/Interperiodica"

gene transfer are not analyzed. Therefore, the approach of Mirkin *et al.* [19] and our approach can be considered complementary.

PRIMARY DATA

The method was applied to the following list containing 40 prokaryotic species from 13 groups:

Archaebacteria: (*Afu*) *Archaeoglobus fulgidus*; (*Hbs*) *Halobacterium* sp. NRC-1; (*Mja*) *Methanococcus jannaschii*; (*Mth*) *Methanobacterium thermoautotrophicum*; (*Tac*) *Thermoplasma acidophilum*; (*Tvo*) *Thermoplasma volcanium*; (*Pho*) *Pyrococcus horikoshii*; (*Pab*) *Pyrococcus abyssi*; (*Ape*) *Aeropyrum pernix*; (*Sso*) *Sulfolobus solfataricus*.

Gamma-proteobacteria: (*Eco*) *Escherichia coli*; (*Buc*) *Buchnera* sp.; (*Pae*) *Pseudomonas aeruginosa*; (*Vch*) *Vibrio cholerae*; (*Hin*) *Haemophilus influenzae*; (*Pmu*) *Pasteurella multocida*; (*Xfa*) *Xylella fastidiosa*.

Beta-proteobacteria: (*Nme*) *Neisseria meningitidis* MC58.

Alpha-proteobacteria: (*Mlo*) *Mesorhizobium loti*; (*Ccr*) *Caulobacter crescentus*; *(Rpr)* *Rickettsia prowazekii*.

Epsilon-proteobacteria: (*Hpy*) *Helicobacter pylori*; (*Cje*) *Campylobacter jejuni*.

Gram-positive bacteria (Firmicutes and Mollicutes): (*Spy*) *Streptococcus pyogenes*; (*Bsu*) *Bacillus subtilis*; (*Bha*) *Bacillus halodurans*; (*Lla*) *Lactococcus lactis*; (*Sau*) *Staphylococcus aureus*; (*Uur*) *Ureaplasma urealyticum*; (*Mpn*) *Mycoplasma pneumoniae*; (*Mge*) *Mycoplasma genitalium*.

Chlamydia: (*Ctr*) *Chlamydia trachomatis*; (*Cpn*) *Chlamydia pneumoniae*.

Spirochetes: (*Tpa*) *Treponema pallidum*; (*Bbu*) *Borrelia burgdorferi*.

Group DMS: (*Dra*) *Deinococcus radiodurans*; (*Mtu*) *Mycobacterium tuberculosis*; (*Syn*) *Synechocystis*.

Thermotoga and Aquifex: (*Aae*) *Aquifex aeolicus*; (*Tma*) *Thermotoga maritima*.

We used the data base COG (clusters of orthologous genes, http://www.ncbi.nlm.nih.gov/COG/). Each COG consists of the genes and proteins encoded by these genes, all having common origin and responsible for a common function. Each COG has multiple alignments of its protein sequences; this allows one to use various methods to obtain a phylogenetic protein tree of this COG. The alignments and the trees were kindly provided by Y. Wolf and E. Koonin (National Center for Biotechnology and Information, USA). The trees were obtained using a combination of the distance method and maximal likelihood method, and this resulted in phylogenetic tree of the genes forming a given COG with distances reflecting similarity of the proteins [15].

Each of the 132 resulting trees has edges with lengths reflecting the supposed evolutionary time between the events ascribed to the terminal nodes of the branch.

Using our algorithm for tree reconciliation [2], we generated the species tree *S* as a tree most similar to these 132 gene trees G_i (where *i* varies from 1 to 132 corresponding to the task 1 below). The resulting species tree is practically identical to the species tree *S** obtained with the fifth method of Wolf *et al.* [16]. The tree *S** was kindly given to us by the authors; this tree was used to identify the events of the horizontal transfer described below.

The graph of the species tree *S** is shown in Fig. 1, and interesting phylogenetic gene trees (for the GOGs discussed in Results and Discussion) are shown in Fig. 2. The table contains the genes selected by our method as candidate horizontally transferred genes.

METHODS

Phylogenetic Trees and Their Reconciliation

Let us consider trees G, $G_1, \ldots, G_n$, S, each with its own number *m* of the terminal nodes. The trees designated by *G* (with or without a subscript) are protein (gene) trees, and the trees designated by *S* (with or without a subscript) are phylogenetic species (cluster, species) trees. Numerous naturally developed ways exist to measure the extent (cost) of the difference between two trees *G* and *S*; the differential cost for the two trees (synonym: *cost* of *G* reconciliation into *S*) is designated $c(G, S)$. Then naturally the difference between a set of trees $G_1, \ldots, G_n$ and tree *S* can be measured as

$$F(G_1, \ldots, G_n, S) = \sum_{i=1}^{n} c(G_i, S).$$

Let us call the functional *F* the *differential cost* (synonym: *degree of reconciliation*) for the given set of trees $G_1, \ldots, G_n$. Let us name the tree *S* that minimizes the *F* a *reconciled tree* (synonym: *consensus tree*).

Let us consider an example to find the differential cost $c(G, S)$ for two trees *G* and *S*. We define *reconciliation* $G \longrightarrow S$ as identical for the terminal nodes and equal to

$$\alpha(x \cup y) = \alpha(x) \cup \alpha(y),$$

where $\cup$ denotes supremum (the smallest precise upper limit of the set $\{\alpha(x), \alpha(y)\}$, i.e., of the set corresponding to the right side of the equation). It is important to distinguish a parent from a supremum, as

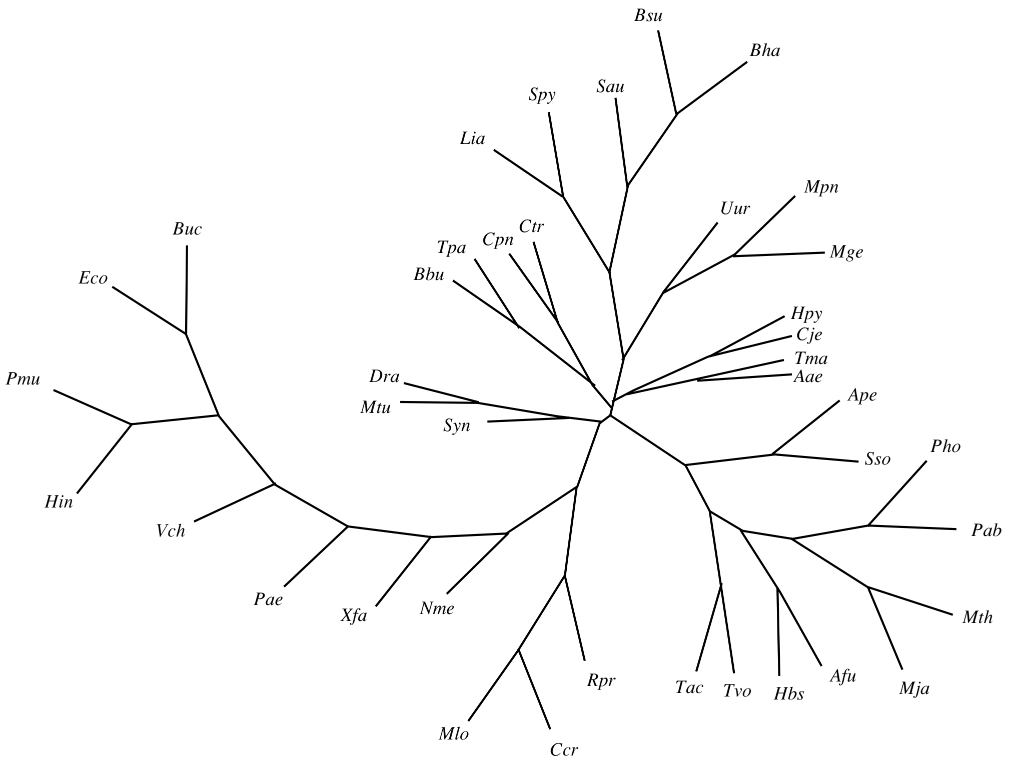
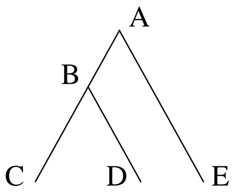


Fig. 1. Phylogenetic tree *S** showing evolution of the 40 studied microbial species.

shown by the example below:



In this case, A is a parent for B and E and a supremum for A, B, C, D, and E; B is a parent for C and D and a supremum for B, C, and D.

More precisely, 'identical' here means the following. If one gene from a species is analyzed, then α is really an identical reflection between the gene and the organism. In general, all genes taken from the organism should be reflected into this organism; if no genes were taken from an organism, then it will reflect no terminal nodes of the tree *G*.

Other definitions of this type of reconciliation are possible taking into account finer models of evolution. It should be noted that development of any, even quite hypothetical models of evolution (producing certain restriction for the background calculation scheme) is interesting for this type of work.

Then $c(G, S)$ is estimated as total penalty for all disorders, either injective (two argument values become one, i.e., when $\alpha(x) = \alpha(y)$ for a pair x, y) or surjective (z value with no *reflection*, i.e., all z for which $z = \alpha(x)$ never becomes true independent of x). In other words, charged are all duplications (one- and two-side) and all losses (intermediate nodes). Various coefficients and parameters can be introduced reflecting one or another model of evolution.

Calculation of Differential Cost for the Two Phylogenetic Trees

In this work we calculated $c(G, S)$ as follows. The pair (g, s), where $\alpha(g) = s$, is named *one-side duplication* if one of the following conditions is true: $\alpha(g) = \alpha(cg)$ or $\alpha(g) = \alpha(c'g)$, where cg is the *left offspring* of node g, and $c'g$ is the *right offspring* of node g. (If both conditions are true, then the pair (g, s) is named *two-side duplication*.) Let us designate $O(G, S)$ the set of all one-side duplications. Node s from tree S is named G-intermediate if it is located strictly between $\alpha(g)$ and $\alpha(pg)$, where pg is *a parent* of node g. Let us designate l_g the set of all g-intermediate nodes, and let us designate $M(S, G)$ the combined set of the sets l_g for all g.

The algorithms designed to generate a phylogenetic protein (gene) tree G sometimes allow ascribing of lengths to the tree edges. We designate these lengths $c(g, g')$, where g and g' are neighboring nodes of the tree G. In some cases these lengths may be interpreted as the time passed between events g and g' within the gene tree G assuming constant evolution rate of the respective gene. We calculated these lengths using

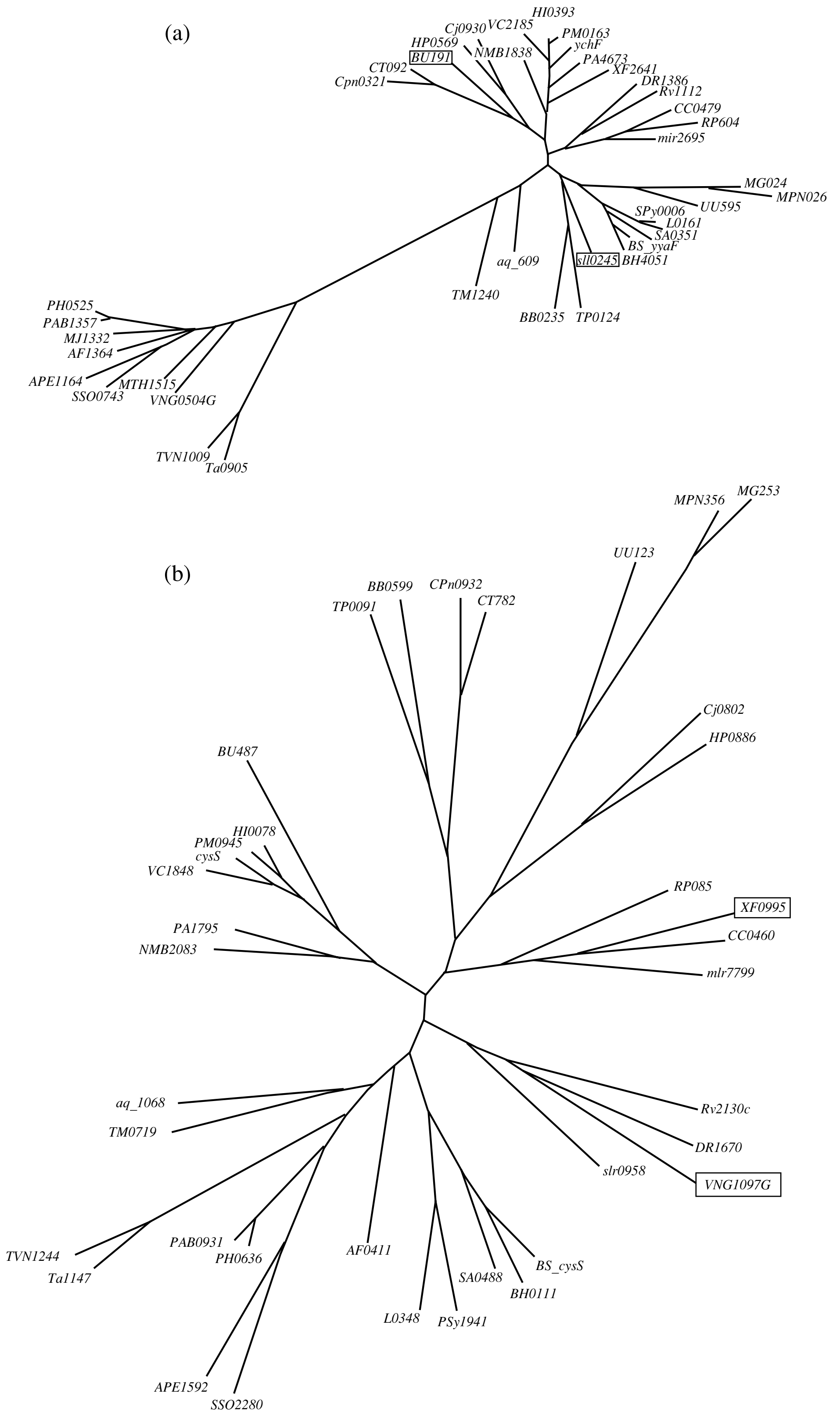


Fig. 2. Examples of horizontal gene transfer. (a) COG0012, predicted GTPase, gene *BU191* from *Buchnera aphidicola*; (b) COG0215, cysteinyl-tRNA synthetase, gene *VNG1095G* from *Halobacterium sp.*; (c) COG0143, methionyl-tRNA synthetase, gene *mlr5926* from *Mesorhizobium loti*; (d) COG0102, ribosomal protein L13, gene *DR0174* from *Deinococcus radiodurans*; (e) COG0198, ribosomal protein L24, gene *BB0489* from *Borrelia burgdorferi*; (f) COG0272, NAD-dependent DNA ligase, gene *yicF* from *Escherichia coli*; (g) COG0343, queuine/archaeosine-tRNA ribosyltransferase, gene *AF1485* from *Archaeoglobus fulgidus*.

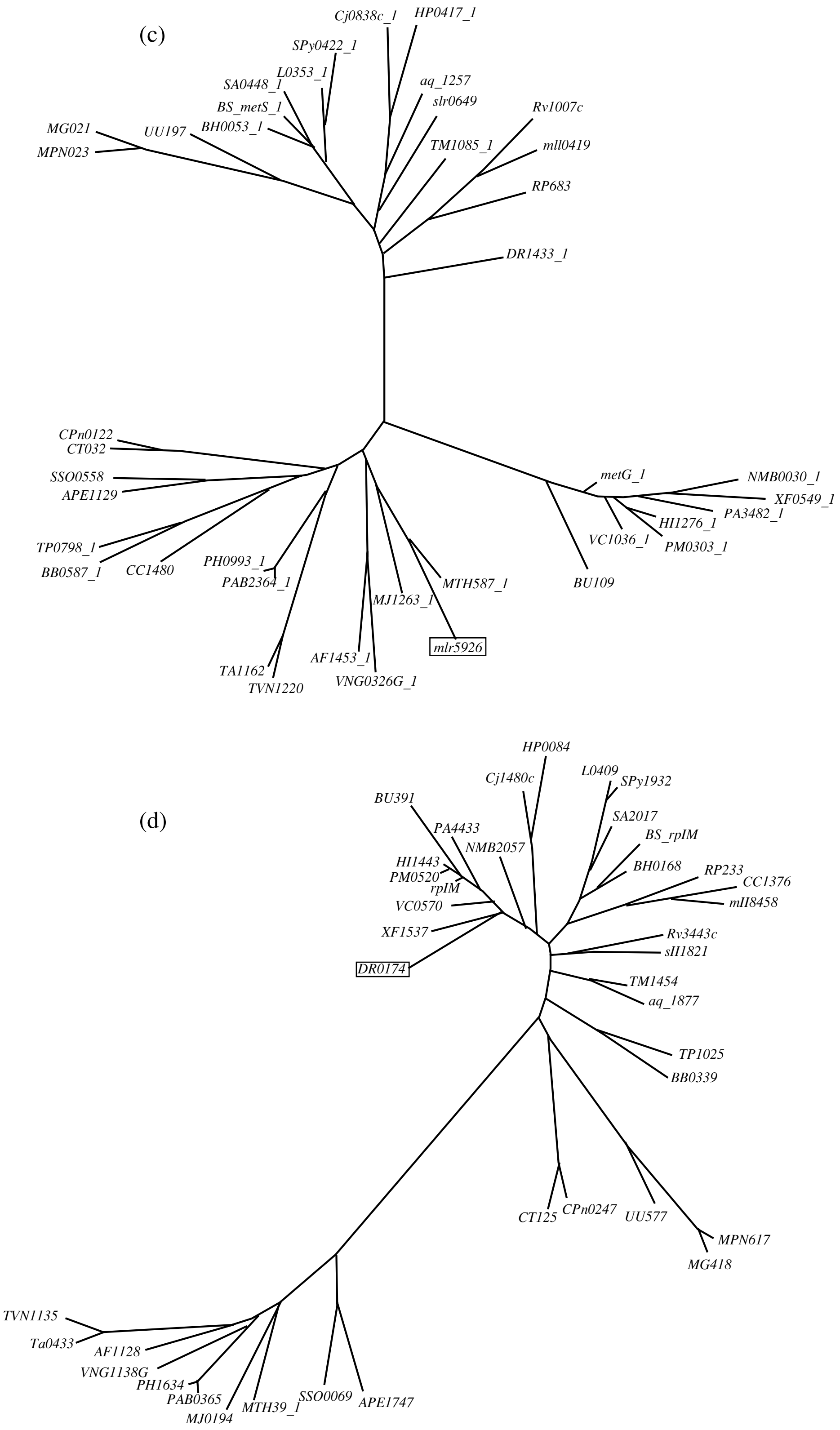

(c)
Cj0838c_1
HP0417_1
SPy0422_1
L0353_1
SA0448_1
aq_1257
slr0649
Rv1007c
BS_metS_1
BH0053_1
MG021
UU197
MPN023
TM1085_1
mll0419
RP683
DR1433_1
CPn0122
CT032
SSO0558
APE1129
TP0798_1
BB0587_1
CC1480
PH0993_1
PAB2364_1
MJ1263_1
MTH587_1
mlr5926
TA1162
TVN1220
AF1453_1
VNG0326G_1
metG_1
NMB0030_1
XF0549_1
PA3482_1
HI1276_1
VC1036_1
PM0303_1
BU109
(d)
HP0084
Cj1480c
L0409
SPy1932
BU391
SA2017
PA4433
BS_rpIM
NMB2057
HI1443
PM0520
rpIM
BH0168
RP233
CC1376
mII8458
VC0570
XF1537
Rv3443c
sII1821
DR0174
TM1454
aq_1877
TP1025
BB0339
CT125
CPn0247
UU577
MPN617
MG418
TVN1135
Ta0433
AF1128
VNG1138G
PH1634
PAB0365
MJ0194
MTH39_1
SSO0069
APE1747

Fig. 2. (Contd.)

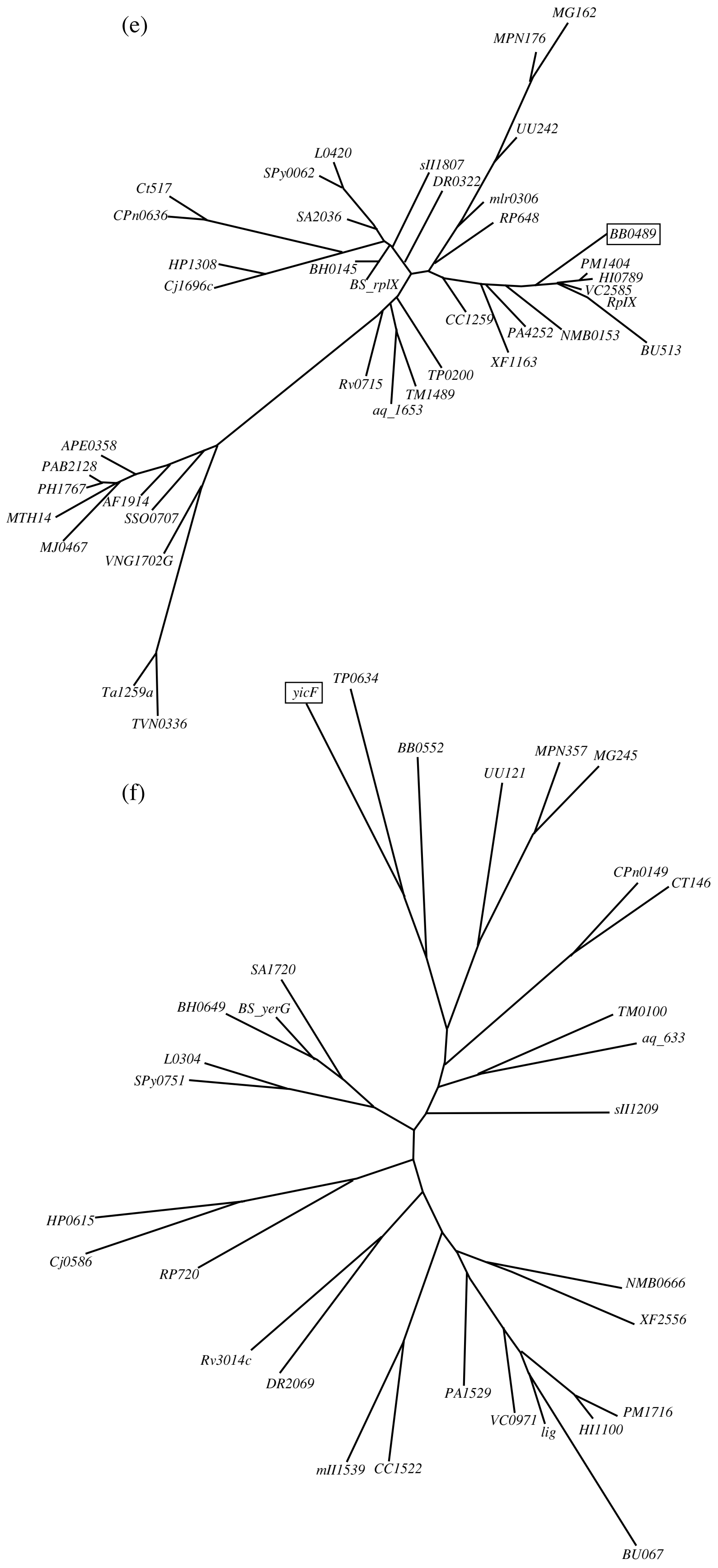

(e)
MG162
MPN176
UU242
L0420
SPy0062
sII1807
DR0322
mlr0306
RP648
Ct517
CPn0636
SA2036
BB0489
HP1308
BH0145
PM1404
HI0789
VC2585
RplX
Cj1696c
BS_rplX
CC1259
PA4252
NMB0153
BU513
XF1163
TP0200
Rv0715
TM1489
aq_1653
APE0358
PAB2128
PH1767
AF1914
MTH14
SSO0707
MJ0467
VNG1702G
Ta1259a
TVN0336
(f)
yicF
TP0634
BB0552
MPN357
MG245
UU121
CPn0149
CT146
SA1720
BH0649
BS_yerG
TM0100
aq_633
L0304
SPy0751
sII1209
HP0615
Cj0586
RP720
NMB0666
XF2556
Rv3014c
DR2069
PA1529
PM1716
VC0971
lig
HI1100
mII1539
CC1522
BU067

Fig. 2. (Contd.)

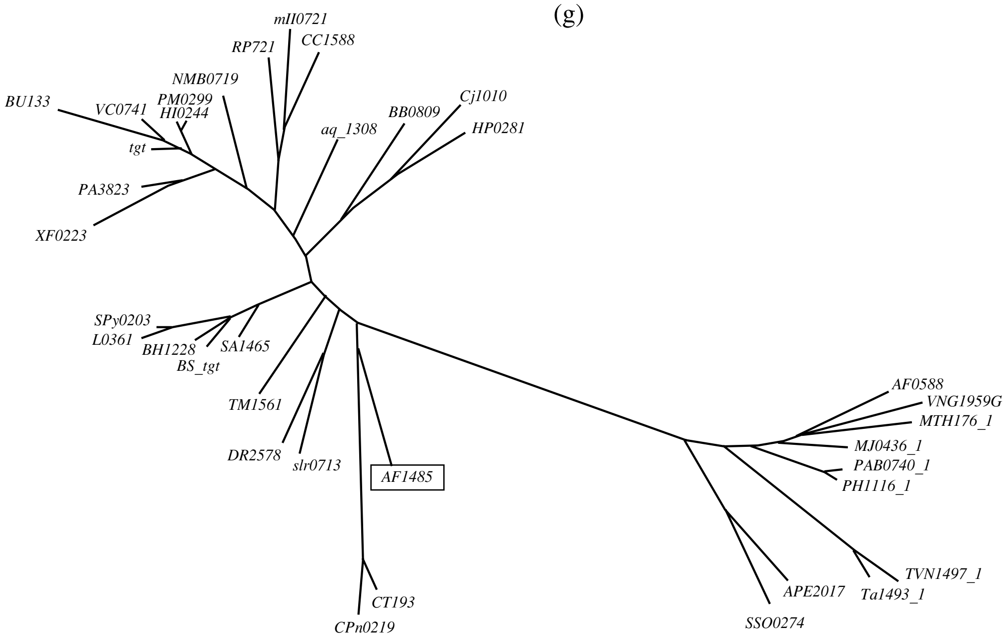


Fig. 2. (Contd.)

either sequence *similarity* (obtained for the nodes *g* and *g*' by the algorithm to generate the tree *G*), or *bootstrap support* for the respective cluster part (part of the tree *G* underlying the edge *gg*'). The calculations below use similarity values. We assumed that

$$c(G, S) = \sum_{(g, \alpha(g)) \in O(G, S)} c(g, pg) + \gamma \sum_{(g, \alpha(g)) \in M(G, S)} c(g, pg)|I_g|.$$

Any unidentified additive was considered zero.

The first member of this sum shows losses from one-side duplication, and the second member shows losses from the missed nodes. The multiplying parameter γ corrects for different frequency of duplications and losses in the course of evolution. Here we assumed $\gamma = 0.1$.

We also considered the third additive of the sum. It characterized quality of the gene tree *G* (e.g., quality of the COG used for the tree mapping); we named it *radiation of the COG*, and it is designated R_G. The results below were obtained without the third additive.

All calculations of the function *F* and value *c* (or similar values) for any gene tree *G* were preceded by *normalization of the lengths* for the tree *G* edges; the aim was to proportionally shorten too long (compared with the mean length) edges of the terminal nodes. Mean length $l_{\text{mean}}(G)$ was calculated for all edges of the terminal nodes from the tree *G*. Then all lengths of the terminal nodes with $l(g) > l_{\text{mean}}$ were modified using the formula:

$$l(g) = (l(g) - l_{\text{mean}})(1 + \mu)^{-(l(g)/l_{\text{mean}})} + l_{\text{mean}},$$

where left part of the equation is new length *l*(*g*) of a long edge, and right part of the equation contains original length *l*(*g*) of the same edge. For all calculations shown here we assumed $\mu = 0.7$. The idea of this normalization is to diminish the impact of the longer edges into the *F* function of the losses aiming to make lower the effect of the reason for the tree incongruence undesirable to consider here, i.e., of the distinct rate of evolution of the same gene family in different phylogenetic lines.

Two Main Tasks

As shown above, function *F* corresponds to $F_{\gamma, \mu}$, therefore we also considered the task to search for parameter values for the function $F_{\gamma, \mu}$ producing within the task 1 (see below) the species tree *S* most similar to the biologically approved (in the given case) species tree *S** (we tried to introduce other parameters related with the evolution, and we also tried to consider variations of *F* in the functional space). This forms an independent task to be discussed elsewhere.

Two main tasks can be considered:

First task. Given the set of gene (protein) trees G_1, …, G_n find the species tree *S* most similar to these trees, i.e., the tree for which differential cost function *F* is minimal either in total or for a given part of the tree.

List of candidate genes involved in evolutionary horizontal transfer events (see text for detailed description of columns)

N	1	2	3	4	5	6	7
1	38. COG0012[*Syn*] *sll0245*	–3.78% (0.1)	2.67 (0.1)	–1.37 –		(3/8 3/8)	1.3
2	41. COG0012[*Buc*] *BU191*	–10.74% (0.02)	3.14 (0.03)	–3.73	Chlamydiae	(3/12 3/12 4/10 4/10)	1.59
3	38. COG0018[*Syn*] *sll0502*	–7.07% (0.07)	2.64 (0.1)	–1.74	Chlamydiae	(3/8 3/8 4/11 4/10)	1.07
4	40. COG0018[*Mtu*] *Rv1292*	–10.74% (0.02)	2.9 (0.05)	–2.62		(2/10 4/10 4/9)	1.38
5	35. COG0061[*Xfa*] *XF2090*	–9.29% (0.03)	1.89 (0.15)	–2.75	alpha-proteobacteria	(2/6 3/6 4/5)	0.93
6	40. COG0072[*Mtu*] *Rv1650_2*	–11.45% (0.03)	2.16 (0.03)	–3.49	alpha-proteobacteria	(3/9 4/9 4/10 4/10 4/3)	1.52
7	41. COG0080[*Bha*] *BH0409*	–6.91% (0.02)	2.17 (0.02)	–2.77	Spirochaetae	(4/10 4/10 4/6)	1.85
8	38. COG0085[*Aae*] *aq_1939*	–15.91% (0.03)	2.33 (0.08)	–3.77	epsilon-proteobacteria	(3/7 3/7)	1.76
9	40. COG0102[*Dra*] *DR0174*	–12.29% (0.03)	2.8 (0.03)	–4.43	gamma-proteobacteria	(2/9 4/11 4/8)	1.73
10	37. COG0126[*Aae*] *aq_118*	–11.59% (0.03)	2.0 (0.05)	–2.87		(3/7 3/7 4/7 4/7)	1.34
11	40. COG0143[*Mtu*] *Rv1007c*	–8.95% (0.05)	3.8 (0.05)	–2.32	alpha-proteobacteria	(2/10 3/9)	1.12
12	41. COG0143[*Mlo*] *mlr5926*	–12.92% (0.02)	4.6 (0.02)	–3.45		(2/11 3/12)	1.53
13	42. COG0162[*Pae*] *PA4138*	–8.12% (0.02)	2.22 (0.05)	–3.05		(2/7 3/7 4/6)	1.47
14	30. COG0173[*Mtu*] *Rv2572c*	–17.47% (0.03)	2.37 (0.03)	–2.89	alpha-proteobacteria	(3/9 4/8 4/10 4/10 4/8)	1.5
15	33. COG0178[*Hbs*] *VNG2636G*	–8.38% (0.03)	2.6 (0.03)	–2.18	DMS	(2/8 4/9 4/9)	1.47
16	30. COG0193[*Dra*] *DR2372*	–16.18% (0.03)	2.67 (0.03)	–2.41	gamma-proteobacteria	(2/7 4/9)	1.26
17	40. COG0198[*Bbu*] *BB0489*	–15.81% (0.03)	2.63 (0.03)	–4.23	gamma-proteobacteria	(3/7 4/12 4/10 4/12 4/9)	1.75
18	38. COG0200[*Bbu*] *BB0497*	–6.64% (0.05)	2.36 (0.08)	–2.75		(3/8 4/9 4/9)	1.55
19	30. COG0203[*Mtu*] *Rv3456c*	–11.92% (0.03)	2.67 (0.03)	–2.73		(3/8 3/8)	1.53
20	38. COG0215[*Xfa*] *XF0995*	–13.86% (0.05)	1.89 (0.05)	–2.5	alpha-proteobacteria	(2/6 3/6 4/5)	1.27
21	39. COG0215[*Hbs*] *VNG1097G*	–20.64% (0.03)	2.5 (0.03)	–3.78	DMS	(2/8 4/9 4/9 4/9)	1.52
22	29. COG0221[*Mlo*] *mlr8562*	–9.12% (0.03)	3.29 (0.1)	–2.24		(3/10 4/13)	1.67

Table. (Contd.)

N	1	2	3	4	5	6	7
23	30. COG0222[*Dra*] *DR2043*	–10.31% (0.03)	3.4 (0.03)	–2.23	alpha-proteobacteria	(2/9 3/8)	1.74
24	35. COG0242[*Syn*] *slr1549*	–10.14% (0.03)	3.2 (0.03)	–2.64	Spirochaetae	(2/8 3/8)	1.47
25	40. COG0250[*Aae*] *aq_1931*	–9.29% (0.02)	2.2 (0.09)	–3.55		(3/7 3/7 4/8)	1.73
26	27. COG0272[*Rpr*] *RP720*	–8.8% (0.1)	1.67 (0.16)	–1.16		(3/5 3/5)	1.4
27	31. COG0272[*Eco*] *yicF*	–29.13% (0.03)	4.8 (0.03)	–4.22	Spirochaetae	(2/12 3/12)	1.64
28	30.COG0292[*Tma*] *TM1592*	–20.24% (0.03)	2.8 (0.03)	–3.04	DMS	(2/7 3/7)	1.76
29	35. COG0294[*Ccr*] *CC3224*	–7.61% (0.06)	3.22 (0.06)	–1.6	DMS	(3/10 3/10 3/9)	0.89
30	30. COG0335[*Dra*] *DR0755*	–6.19% (0.07)	3.6 (0.03)	–1.7 –	Chlamidiae	(2/9 3/9)	1.77
31	36. COG0343[*Afu*] *AF1485*	–8.58% (0.06)	3.33 (0.03)	–2.2	Chlamydiae	(3/10 3/10)	1.6
32	30. COG0359[*Aae*] *aq_2042*	–13.19% (0.03)	2.8 (0.03)	–2.64	DMS	(2/7 3/7)	1.39
33	40. COG0441[*Ape*] *APE0809*	–14.09% (0.02)	3.0 (0.07)	–3.25	DMS	(2/8 3/7)	1.79
34	29. COG0452[*Bbu*] *BB0812*	–8.69% (0.03)	2.33 (0.03)	–2.34		(2/7 4/7)	1.35
35	37. COG0504[*Tpa*] *TP0305*	–21.5% (0.03)	3.2 (0.05)	–3.95	DMS	(2/8 3/8)	1.64
36	40. COG0525[*Rpr*] *RP687*	–16.16% (0.03)	3.67 (0.03)	–3.99		(3/12 3/12 3/9)	1.62
37	33. COG0547[*Cje*] *Cj0346_2*	–12.66% (0.03)	3.6 (0.09)	–2.6		(2/9 3/9)	1.28
38	39. COG0552[*Dra*] *DR2260*	–13.72% (0.03)	2.27 (0.05)	–3.14	alpha-proteobacteria	(3/8 4/9 4/9 4/8)	1.52
39	29. COG0556[*Hbs*] *VNG2390G*	–9.23% (0.03)	2.33 (0.03)	–1.93 –	DMS	(3/9 3/8 3/4)	1.53
40	28. COG0571[*Syn*] *slr0346*	–19.72% (0.07)	2.67 (0.1)	–2.43	Chlamydiae	(3/8 3/8 3/8)	1.56
41	35. COG0587[*Bsu*] *BS_yorL*	–10.28% (0.03)	3.11 (0.03)	–2.5		(3/10 3/10 3/8)	1.6

Second task. Given the protein (gene) tree G and the species tree S with high differential cost $c(G, S)$, find a terminal node in G (i.e., gene or protein) or a group of nodes responsible for the high cost. This means, after removal of this node(s) the differential cost of the resulting trees G' and S' should become low.

Thereby, we intent to identify the candidate genes involved in horizontal transfer events in the course of their evolutionary history. It is also important to estimate quality of the COGs (see [2]).

Only the second task is considered in this work.

Gene *g* (a terminal node within the gene tree *G*) is determined as a candidate for a horizontal transfer event with the following procedure. Taxonomy is usually fixed in the species tree, and it can be either one- or multilevel, means that it is represented by a certain graph with groups of species as nodes.

The neighborhood of radius *r* centered in terminal node *g* is defined as including all terminal nodes *g*' within the tree *G* with distance from g bigger than or equal to *r*; distance between two nodes in tree *G* is defined as number of edges on the shortest way between them within *G*. The neighborhood of radius *r* with the center in *g* (when *g* is a terminal node of the gene tree *G*) without the central point *g* is named *centerless*. If reconciliation α reflects the centerless neighborhood into one or several neighboring species groups, while center *g* itself goes into a species distant from these groups (the distance is taken from the species tree *S*), then *g* is considered a candidate for a horizontal transfer *event* (within this criterion). Search for these genes implies calculation and comparison of several parameters of incongruence *statistics* developed basing on the idea explained above.

We considered the groups of organisms listed above in the Primary data section as equal taxonomic units; i.e., in *S** we used the *one-level taxononomy.*

Any group or set of groups within the species tree *S** has exactly one node with subtree (cluster) covering the joint set of these groups; these subtree is minimal among all possible subtrees. This subtree is named *conditional group* (of species or organisms) within *S**, and the respective node is called *the root* of this group. Obviously, if formed by members of the same taxonomic group, the conditional group falls within a single taxonomic group. We define the size of the conditional group as the maximal distance (sometimes as statistically significant distance) from its root to its terminal node. We define the distance between any conditional group and any node of the tree *S** as the number of edges on the shortest way from the group root and the node; this distance is taken positive if the node falls into the group, or negative if not.

Thereby, the above definition of horizontal transfer indicates that the reconciled centerless neighborhood of the gene *g* should form a small conditional group with big distance to reflection of the gene *g* itself.

We have also analyzed a more complicated multilevel taxonomy in the tree *S**, these results will be published elsewhere.

Statistics to Identify Candidate Genes Involved in Horizontal Transfer

Let us define statistics for the second of the abovementioned tasks. The results of calculations for some genes are shown in the table, and complete calculation results for the whole set of the studied genes are available online at http://www.iitp.ru/lyubetsky. Below we explain the statistics shown in the columns of the table, column by column.

The leftmost column contains numbers for the rows. **Column 1** shows gene number within the COG, the COG number, abbreviated name of the species, and name of the gene.

Column 2 contains the results obtained for the relative increase in reconciliation cost F_g for the gene *g* for the trees *G* and *S*. The percentage was calculated as:

$$F_g = ((c_g - c)/c) \times 100,$$

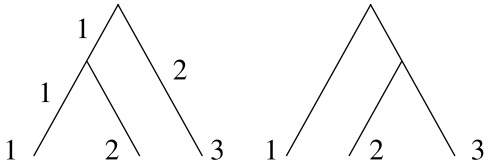
where *c* is the cost of reconciling *G* into *S* and c_g is the cost of reconciling G_g into *S*. Here G_g is generated from *G* by deletion of the terminal node *g*. When a terminal node is deleted, new edges appear with lengths equal to the total length of the “disappeared pathway.” Parenthesized in column 2 is the *p* value calculated as follows:

$$p(g) = card(\{g': F_{g'} \le F_g\})/m,$$

where *card*(*X*) is the number of elements in the set *X*, and *m* is the number of terminal nodes (genes) in the tree *G*. The program is designed to select and to type out the genes from *G* for which F_g is low enough considering critical values P_0 and $p(g) \le p_0$, where p_0 is any given appropriate critical value.

In other words, for a rather low p_0 we selected all cases of extreme position for the gene *g* within the gene tree, i.e., all genes *g* for which F_g is extremely low compared with most other values of $F_{g'}$. For this table we normally took $P_0 = -7$ and $p_0 = 0.1$.

Example 2A.

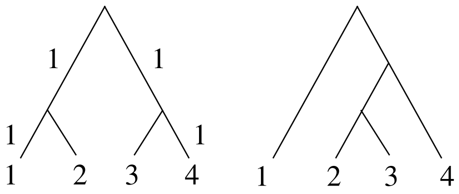


The gene tree *G* (left) has equal lengths of 1 of the edges for nodes 1 and 2. The cost *c* of its reconciliation into species tree *S* (right) is 1.16 (assuming $\gamma = 0.1$, $\mu = 0.7$).

G	1	2	3
F_g (no normalization)	–100%	–83.3%	–83.3%
F_g (normalized)	–100%	–82.8%	–82.8%

The genes of the tree *G* have become ordered; gene 1 is stronger candidate, while genes 2 and 3 are equal in this respect.

Example 2B.



The gene tree *G* (left) has all edges equal to 1. The cost *c* of its reconciliation into species tree *S* (right) after normalization of *G* is 1.3 (assuming $\gamma = 0.1$, $\mu = 0.7$).

G	1	2	3	4
F_g (no normalization)	0%	–92.3%	7.7%	23.1%
F_g (normalized)	–2.85%	–92.3%	4.85%	14.5%

The genes of the tree *G* have become ordered; gene 2 is the strongest candidate.

Column 3 of the table shows the reversed ratio of the mean distance between the gene *g* and all his neighbors *g*' from the area of radius *r* on the gene tree *G* to the mean distance between α(*g*) and α(*g*') on the species tree *S* (*r* in this table equals 4). We designate this ratio R_g and name it radiation of the gene *g*. It is calculated as follows:

$$R_g = \frac{\left(\sum_{g'} \rho(\alpha(g), \alpha(g'))\right)/(m'_g - 1)}{\left(\sum_{g'} \rho(g, g')\right)/(m_g - 1)},$$

where m_g is number of elements in this area and m'_g is number of elements in reflection of this area. Note that the distance ρ between *g* and *g*' is calculated as number of edges on the shortest way between *g* and *g*' in *G* (or between α(*g*) and α(*g*') in *S*).

The *p* value parenthesized in column 3 is calculated similarly to the *p* value in column 2 as:

$$p(g) = card(\{g': R_{g'} \geq R_g\})/m.$$

In this case the candidate genes for horizontal transfer are selected as all genes *G* with extremely high R_g compared with most other R_g values. Beside, the following condition was considered essential.

If phylogenetic gene tree neighbors of the gene *g* in species tree are located far from the gene *g* carrier, we assume that gene *g* could be transferred to the ancestor of the α(*g*) species from the ancestors of other species. This possibility grows stronger when the species containing *g*-related genes belong to one and the same taxonomic group. The candidate species or groups being a putative source for horizontal transfer of the gene *g* are listed in column 5.

In other words, this statistics selected genes *G*, for which R_g is rather high compared with critical P_0 and *p*(*g*) and rather low compared with critical p_0 (in these calculation we usually assumed $P_0 = 2$ and $p_0 = 0.1$), and the above condition is true. Interestingly, the empirical distribution of statistics R_g for all analyzed COGs is similar to one and the same standard distribution

Example 3. For the trees from example 2B and *r* = 3.

G	1	2	3	4
R_g	2	2	1.5	1.5

The genes of the tree *G* have become ordered; genes 1 and 2 are stronger candidates than genes 3 and 4.

Column 4 of the table shows deviation *var*(*g*) (as standard deviation), i.e., from respective increase of reconciliation cost F_g for the a given *g* we subtract mean value $F_{\overline{g'}}$ of the $F_{g'}$ for all terminal nodes *g*' of the given gene tree *G*, and divide the results by σ.

$$var(g) = \frac{F_g - (F_{\overline{g'}})}{\sigma},$$

$$\sigma = \sqrt{(1/m - 1)\sum_g (F_g - (F_{\overline{g'}}))^2}.$$

Here σ is calculated using a common formula, where *m* is the number of elements in the given COG tree.

For our primary data, empirical distribution of statistics F_g is close to normal distribution. Therefore, we apply a statistically reliable procedure selecting the genes for which the absolute value of *var*(*g*) is higher than 2. The possibility of this event is 0.05; therefore, the respective genes may be considered as those of extreme location. Minus sign in column 4 means that condition $|var(g)| \geq 2$ is not true.

The statistics shown in columns 2 and 4 of the table provide similar results of candidate gene selection if the distribution of statistics F_g is normal. Statistics of column 2 is more universal since it does not depend on distribution of F_g, while statistics of column 4 has better justification.

Example 4. Using the data from example 2B we have $(\bar{F}_{g'}) = -18.95$, $\sigma = 49.4$.

G	1	2	3	4
$Var(g)$	+0.33	–1.485	0.48	0.68

The genes of the tree G have become ordered; only gene 2 can be considered candidate.

Column 5 of the table shows a group considered a putative source of the horizontally transferred gene g. (Here $r = 4$, other values of r can be tried including those big enough for the neighborhood to cover the set of all terminal nodes; we talk about the centerless neighborhood of radius r centered in node g). The shown group can be either taxonomical or conditional group defined by reflection of the given centerless neighborhood.

Column 6 of the table lists the relative increment in reconciliation cost $F_{g'}$ for the neighbors g' of the gene g, as

$$\rho(g, g')/\rho(\alpha(g), \alpha(g')),$$

where the numerator is distance from g to its neighbor within the tree G, and the denominator is the distance from $\alpha(g)$ to $\alpha(g')$ in the species tree S.

Example 5 shows calculations of $F_{g'}$ in case when gene g' is lost, while gene g is retained. The values of $F_{g'}$ are not shown in the table; this information is considered complementary.

Example 5. Data from example 2B and $r = 3$.

G	1	2	3	4
$Var(g)$	$2/4 : F_2 = -92.3$	$2/4 : F_1 = -2.85$	$2/3 : F_4 = 14.5$	$2/3 : F_3 = 4.85$

Column 7 of the table shows the ratio of the edge length for the terminal node (gene) g after normalization (see example 2) and the mean edge length for the terminal node (gene) g for all genes g' of the given tree G, calculated as $l(g)/l_{\text{mean}}(G)$. This ratio may be considered a measure of occurrence of the long edges for the whole tree and for its terminal nodes.

To summarize, the algorithm select as candidate genes for horizontal transfer events those genes which follow the above rules using the statistics listed in columns 1–4 of the table; columns 5 and 6 of the table are complementary and contain information that can prove useful for an expert.

Obviously, selection of the critical values discussed above and interpretation of the information from the table cannot be considered a formal computer-assisted procedure. At the present step of research, one could hardly expect strict computer procedures to identify horizontal transfer events. Rather, we aimed to find statistical parameters to be included in a more formal procedure to be developed in future.

Finally, the COGs in the table are listed by increasing COG number, and the genes from the same COG are listed following their left-to-right arrangement within the respective gene tree G.

It should be noted that for certain aims of biological analysis the table can be rearranged, sorting its data by the genomes (i.e., by the species) affected by horizontal gene transfer events, and sometimes by taxonomic groups found to be sources of horizontally transferred genes. The complete table including all studied genes is available online (http://www.iitp.ru/lyubetsky).

RESULTS AND DISCUSSION

Some interesting examples are shown in Fig. 2.

Two candidate horizontally transferred genes are readily identified by our method in COG0012 (hypothetic GTPase) (Fig. 2a). An aphid endosymbyont *Buchnera aphidicola* is a gamma-proteobacterium closely related to *E. coli*, however its gene *BU191* on the gene tree is clustered with the genes of *Chlamydia*, the latter being the putative source for the horizontally transferred gene. Similarly, gene *Sll0245* has probably been introduced into the genome of cyanobacterium *Synechocystis* sp. from Spirochetae.

In COG0215 (cysteinyl-tRNA synthetase) gene *VNG1095* from halophylic archaebacterium *Halobacterium* sp. (Fig. 2b) is of eubacterial origin. Its putative source is a genome related to that of *Deinococcus radiodurans*. Within the same cluster we cannot exclude horizontal transfer of the gene *XF0995* from *Caulobacter crescentus*-related alpha-proteobacterium into gamma-proteobacterium *Xylella fastidiosa*. Horizontal transfer in the opposite direction, i.e., from Archaebacteria to Eubacteria was detected in COG0143 (methionyl-tRNA synthetase) (Fig. 2c), where gene *mlr5926* of the alpha-proteobacterium *Mesorhizobium loti* originates from a methanogenic archaebacteriumn and probably was horizontally transferred from Archaebacteria to Eubacteria. Interestingly, this event resulted in paralogous genes in the genome of *M. loti*, since the original gene *mll0419* was retained. It has now become evident, contrary to the earlier assumptions, that the genes encoding aminoacyl-tRNA synthetases often have complex evolutionary history including events of duplication and horizontal transfer [12].

In two cases we detected horizontal transfer of the genes encoding ribosomal proteins (Fig. 2d,e). Gene *DR0174* (Fig. 2d) from the genome of *Deinococcus radiodurans* encodes ribosomal protein L13 and is of gamma-proteobacterial origin, and gene *BB0489* (Fig. 2e) encodes ribosomal protein L24 has been horizontally transferred from either beta- or gammaproteobacteria. Duplication and horizontal transfer events in evolutionary history of the ribosomal proteins was shown earlier [17].

Finally, the *E. coli* gene *yicF*, which codes for NAD-dependent DNA ligase (COG0271), has been transferred from genome of a spirochete; the *E. coli* genome also retains its native gene *lig* (Fig. 2f). The gene encoding queuine/archaeosine-tRNA ribosyltransferase *AF1485* from *Archaeoglobus fulgidus* (COG0343) is probably of eukaryotic origin (Fig. 2g).

To summarize, both large-scale analysis and the study of particular cases show practical applicability of our method.

Obviously, it is practically impossible to experimentally confirm the facts of horizontal transfer, this being true for any evolutionary event. It should be also noted that the difference between the gene tree and the species tree can be explained, beside evolutionary events (duplications, deletions, horizontal gene transfer) considered here and earlier [2] by the drawbacks of the procedure used to generate a tree. These procedures may be inadequate to the evolution process. For the gene tree, errors can arise if gene evolution has different rates in different branches, and for the species tree, distant nodes may have low significance.

Expert evaluation often coincides with the results obtained using our computer-assisted procedures. However, expert analysis of many hundreds of evolution trees is impossible; therefore, one of the main applications of the proposed approach could be the large-scale analysis of protein families aiming to identify candidate cases of the gene transfer for further scrutiny.

Though the results presented here are somewhat preliminary, some interesting conclusions can already be drawn. First, the classes of orthologs with detected horizontal transfer events are mainly formed by the genes that code for the proteins involved in the main cell information processes, most commonly in translation. Probably this is not related to a specific increase of horizontal transfer events in this type of the genes, but rather to the clear and well-resolved structure of the respective phylogenetic trees, allowing one to clearly identify horizontal transfer events. Identification of the horizontal transfer events in families of transporters or transcription regulators requires more effort and more careful analysis.

Second, horizontal transfer is often not accompanied by loss of the original native gene. This may be explained by either an intermediate state when the choice between the two genes of the same function has not yet been fixed, or by some difference in function of the respective protein products. This can be checked experimentally, using successive inactivation of the two variants and/or artificial horizontal transfer events introducing a new gene in the genome with and without inactivation of the original native gene.

This work is based on our earlier publications [1, 2, 18].

ACKNOWLEDGMENTS

The authors are very grateful to Y. Wolf who provided the initial data set for calculations, to E. Koonin for his advice, collaboration and helpful discussion, and to an anonymous reviewer for his interesting and valuable comments.

This work was supported by the Howard Hughes Medical Institute (HHMI grant 5000309) and by Ludwig Institute for Cancer Research (LLCR-CRDF, grant RB0-1268).

REFERENCES

1. V’yugin, V.V., Gelfand, M.S., and Lyubetsky, V.A. 2002. Tree Reconciliation: Reconstruction of Species Phylogeny by Phylogenetic Gene Trees. *Mol. Biol.*, **36**, 650–658.
2. V’yugin, V.V. and Lyubetsky, V.A. 2001. On an algorithm to search for horizontal gene transfer events basing on phylogenetic protein trees. *Informatsionnye Protsessy,* **1**, 167–177.
3. Logsdon J.M., Faguy D.M. 1999. Thermotoga heats up lateral gene transfer. *Curr. Biol.* **9**, 747–751.
4. Bacterial Conjugation. 1993. Ed. Clewell D.B., New York: Plenum Press.
5. Bergh O., Borsheim K.Y., Bratbak G., Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature*. **340**, 467–468.
6. Boucher Y., Doolitle W.F. 2000. The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol. Microbiol*. **37**, 703–716.
7. Lawrence J.G. 1997. Selfish operons and speciation by gene transfer. *Trends Microb.* **5**, 355–359.
8. Lawrence J.G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol*. **2**, 519–523.
9. Doolitle W.F. 1999. Lateral Genomics. *Trends Cell. Biol*. **9**, 5–8.
10. Lawrence J.G., Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA*. **95**, 9413–9417.
11. Nelson K.E., Clayton R.A., Gill S.R. *et al.* 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thetmotoga maritima*. *Nature*. **399**, 323–329.
12. Yanai I., Wolf Y., Koonin E. 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*. **3**, Research 0024.
13. Koonin E.V., Makarova K.S., Aravind L. 2001. Horizontal gene transfer in prokaryotes: quntification and classification. *Annu. Review Microbiol*. **55**, 709–742.
14. Page R.D.M., Charlstone M.A. 1998. From gene to organismal phylogeny: reconciled trees and gene tree/species tree problem. *Mol. Phyl. Evol*. **7**, 231–240.

15. Page R.D.M. 1998. Genetree: comparing gene and species phylogenies using reconciled trees. *Bioinform. Appl. Notes*. **14**, 819–820.

16. Wolf Y., Rogozin I., Grishin N., Tatusov R., Koonin E. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol*. **1**, 8.

17. Makarova K.S., Ponomarev V.A., Koonin E.V. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol*. **2**(9), Research 0033.

18. Lyubetsky V.A., V'yugin V.V. 2002. Method of horizontal gene transfer determination using phylogenetic data. *Proc. Third Internat. Conf. Bioinform. Genome Regulat. Struct.* **2**, IC&G. Novosibirsk, 60–62.

19. Mirkin B.G., Fenner T.I., Galperin M.Y., Koonin E.V. 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 eukaryotes. *BMC Evol. Biol.* **3**, 2.