30 Kaliman, P. *et al.* (1999) Insulin-like growth factor-II, phosphatidylinositol 3-kinase, nuclear factor-kappaB and inducible nitric-oxide synthase define a common myogenic signaling pathway. *J. Biol. Chem.* 274, 17437–17444
31 Corvera, S. and Czech, M.P. (1998) Direct targets of phosphoinositide 3-kinase products in membrane traffic and signal transduction. *Trends Cell Biol.* 8, 442–446
32 Walker, E.H. *et al.* (1999) Structural insights into phosphoinositide 3-kinase catalysis and signalling. *Nature* 402, 313–320
33 Hooshmand-Rad, R. *et al.* (2000) The PI 3-kinase isoforms p110(alpha) and p110(beta) have differential roles in PDGF- and insulin-mediated signaling. *J. Cell Sci.* 113, 207–214
34 Inukai, K. *et al.* (1997) p85alpha gene generates three isoforms of regulatory subunit for phosphatidylinositol 3-kinase (PI 3-kinase), p50alpha, p55alpha, and p85alpha, with different PI 3-kinase activity elevating responses to insulin. *J. Biol. Chem.* 272, 7873–7882
35 Jackson, T.R. *et al.* (2000) Cytohesins and centaurins: mediators of PI 3-kinase-regulated Arf signalling. *Trends Biochem. Sci.* 25, 489–495

T. Cowen
Dept of Anatomy and Developmental Biology, Royal Free and University College Medical School, Royal Free Campus, Rowland Hill St, London, UK NW3 2PF.
e-mail: tcowen@rfc.ucl.ac.uk

Genome Analysis

How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies

Tamas Rujan and William Martin

It is well known that chloroplasts and mitochondria donated many genes to nuclear chromosomes during evolution – but how many is 'many'? A sample of 3961 *Arabidopsis* nuclear protein-coding genes was compared with the complete set of proteins from yeast and 17 reference prokaryotic genomes, including one cyanobacterium (the lineage from which plastids arose). The analysis of 386 phylogenetic trees distilled from these data suggests that between ~400 (1.6%) and ~2200 (9.2%) of *Arabidopsis* nuclear genes stem from cyanobacteria. The degree of conservation preserved in protein sequences in addition to lateral gene transfer between free-living prokaryotes pose substantial challenges to genome phylogenetics.

Chloroplasts are descendants of free-living cyanobacteria, but they have highly reduced genomes. Higher plant chloroplast genomes encode about 80 proteins, the more diverse plastids among algae encode anywhere between 60 and 200 proteins, and non-photosynthetic plastids encode as few as 23 proteins[1]. This is in contrast to cyanobacteria, *Synechocystis*[2] for example, with over 3000 protein-coding genes. Despite this genome reduction, plastids seem to contain a similar number of proteins as cyanobacteria[1,2]. The vast majority of these proteins are encoded by nuclear genes – many of these originally transferred to the nucleus from the cyanobacterial symbiont – although some are still encoded among plastid genomes[3]. A recent estimate based on three-way BLAST[4] comparisons of *Arabidopsis* proteins to homologues from *Synechocystis* and from yeast (as a comparison for genes existing in eukaryotes before the acquisition of plastids), led to an estimate that about 1400 (5.8%) of the *Arabidopsis* genome might have been acquired from the ancestral cyanobacterial symbiont[5]. BLAST analyses are occasionally used as an estimator of phylogeny and lateral gene transfer[6]. But corresponding estimates based on large-scale protein phylogeny are desirable for comparison because BLAST results merely summarize the similarity of one sequence to many others, whereas phylogenetic trees summarize the similarities of many sequences simultaneously.

Here we examine 368 phylogenetic trees constructed from several thousand protein-coding genes and 18 reference genomes to estimate the fraction of *Arabidopsis* genes that come from cyanobacteria. Our criterion for scoring an *Arabidopsis* gene as originating from cyanobacteria is simple – we ask 'is the *Arabidopsis* protein more similar to its cyanobacterial homologue than it is to homologues from any other reference genome?' Translated into the language of phylogenetic trees, that means asking 'does the *Arabidopsis* protein share a common branch with its cyanobacterial homologue in a protein phylogeny?' Using that criterion in protein maximum-likelihood trees and using simple statistical tests, we find that between 1.6% and 9.2% – a large margin of uncertainty – of the *Arabidopsis* genes examined are likely to be acquisitions from cyanobacteria. We find evidence to suggest that the differing levels of sequence conservation among proteins might be responsible for this margin of uncertainty. Furthermore, we outline some of the premises involved in inferring eukaryotic gene origins and how lateral gene transfer between free-living prokaryotes complicates that issue.

Automated sequence filtering for automated phylogenetic analysis

How does one obtain and evaluate a large number of protein phylogenies starting from 3961 proteins and 18 reference genomes using automated procedures? Our strategy is outlined in Fig. 1. (By the time this article appears, the complete *Arabidopsis* genome will have been published, but those data were not available when we embarked upon this work.) We obtained 3961 annotated *Arabidopsis* nuclear-encoded nonredundant proteins (kindly provided by H-W. Mewes, MIPS, Munich). All of the proteins from 17 sequenced prokaryotic genomes plus yeast were retrieved and assembled into a single file before BLAST searching so that the expectations for hits in this mini-database are independent of variations in genome size. The *Arabidopsis* proteins were compared with that set using BLAST.

The BLAST results (tables) from the 3961 *Arabidopsis* proteins were analysed for the appearance of a *Synechocystis* homologue in the table with an

0168–9525/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02209-5

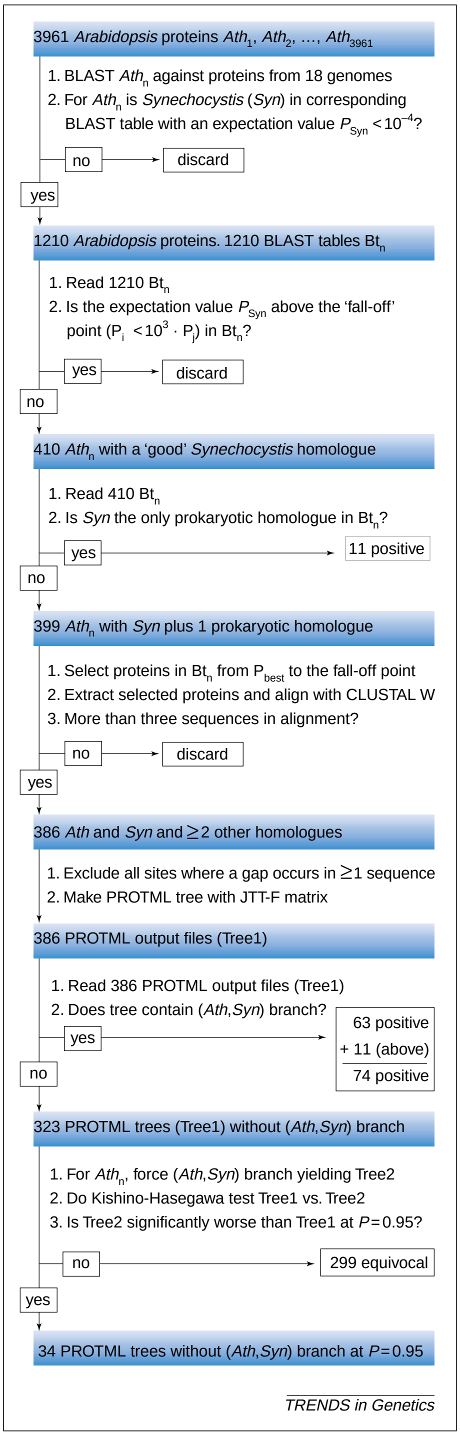

3961 Arabidopsis proteins Ath_1, Ath_2, …, Ath_{3961}
1. BLAST Ath_n against proteins from 18 genomes
2. For Ath_n is Synechocystis (Syn) in corresponding BLAST table with an expectation value $P_{Syn} < 10^{-4}$?
no
discard
yes
1210 Arabidopsis proteins. 1210 BLAST tables Bt_n
1. Read 1210 Bt_n
2. Is the expectation value P_{Syn} above the 'fall-off' point ($P_i < 10^3 \cdot P_j$) in Bt_n?
yes
discard
no
410 Ath_n with a 'good' Synechocystis homologue
1. Read 410 Bt_n
2. Is Syn the only prokaryotic homologue in Bt_n?
yes
11 positive
no
399 Ath_n with Syn plus 1 prokaryotic homologue
1. Select proteins in Bt_n from P_{best} to the fall-off point
2. Extract selected proteins and align with CLUSTAL W
3. More than three sequences in alignment?
no
discard
yes
386 Ath and Syn and ≥2 other homologues
1. Exclude all sites where a gap occurs in ≥1 sequence
2. Make PROTML tree with JTT-F matrix
386 PROTML output files (Tree1)
1. Read 386 PROTML output files (Tree1)
2. Does tree contain (Ath,Syn) branch?
yes
63 positive
+ 11 (above)
74 positive
no
323 PROTML trees (Tree1) without (Ath,Syn) branch
1. For Ath_n, force (Ath,Syn) branch yielding Tree2
2. Do Kishino-Hasegawa test Tree1 vs. Tree2
3. Is Tree2 significantly worse than Tree1 at $P=0.95$?
no
299 equivocal
yes
34 PROTML trees without (Ath,Syn) branch at $P=0.95$
TRENDS in Genetics

Fig. 1. Filtering trees from genomes. A flow diagram summarizes our procedures to identify, select, align and phylogenetically analyse the protein data.

expectation value (the probability of finding a sequence-similarity match of a given quality in a database of a given size purely by chance[4]) $e < 10^{-4}$, leaving 1210 proteins. Those 1210 BLAST tables were then re-examined to see whether the *Synechocystis* homologue was among the 'good' hits; that is, to see whether the *Synechocystis* homologue was among those proteins that are sufficiently similar to the *Arabidopsis* protein using the BLAST search criterion to justify even looking for a common branch for the *Synechocystis* and *Arabidopsis* proteins in later phylogenetic analysis (see below).

But to determine what 'good' is, we had to introduce a criterion that would allow our computer to read the BLAST tables so as to see where the 'good' hits end and the 'poorer' ones begin. We call this criterion the fall-off point. It uses the circumstance that hits in a BLAST table are listed from top to bottom in order of descending quality. The fall-off point is determined by reading the expectation values from the best hit downwards in the BLAST table, taking the quotient of the two adjacent expectation values along the way (i.e. in a 1/2, 2/3, 3/4, 4/5, ... i/j manner) and noting the corresponding organism names in the process. The fall-off is the point at which the quotient reaches a specified threshold, for example 10^3. As a threshold, we found 10^3 (shown as $P_i < 10^3 \cdot P_j$ in Fig. 1) to be convenient, because larger thresholds simply included sequences that are more distantly related to the *Arabidopsis* query. If *Synechocystis* was above the fall-off point, it was considered to be among the 'good' hits; if it was below the fall-off, the corresponding *Arabidopsis* protein and BLAST table protein was discarded.

Using the fall-off-point criterion, we identified 410 *Arabidopsis* proteins that have a 'good' *Synechocystis* homologue, thus warranting further investigation through phylogenetic analysis. In terms of gene families, these 410 encompass 138 unique proteins and 60 *Arabidopsis* gene families ranging from two to 16 members, the exception being the family of serine/threonine protein kinases, that encompasses 79 members. Each member of each gene family was analysed individually as a potential candidate for a gene of cyanobacterial origin.

For 11 of the 410 proteins, *Synechocystis* was the only homologue detected in the data. These 11 genes are shared by *Arabidopsis* and *Synechocystis* but no other genome sampled, and were thus scored as genes that were acquired from plastids (Fig. 1). Subtracting these leaves 399 proteins. From those 399 BLAST tables, the 'good' homologues were selected for alignment and phylogenetic analysis. Here, as above, 'good' sequences were those above the fall-off point. This selection procedure thus gathers sequences in order of decreasing similarity (as assessed by BLAST) to the *Arabidopsis* query and stops gathering when the increment in similarity from one sequence to the next drops off abruptly. If an organism produced more than one match per BLAST table, only the best one from that genome was taken for alignment.

The proteins thus selected were written out into files and aligned using CLUSTAL W (Ref. 7). Thirteen of the alignments so obtained contained only three sequences and were not considered further (there is not much point in making a tree of three sequences) leaving 386 proteins for phylogenetic analysis (Fig. 1). Notably, many of the alignments had poorly conserved regions that contain numerous gaps. To avoid obtaining spurious results, which such highly gapped regions might produce, positions in the alignment at which a gap existed in ≥1 sequences were excluded from analysis (purged) using the corresponding tools provided in PROTML (Ref. 8; available at http://www.ism.ac.jp/software/ismlib/softother.e.html#molphy). Trees were then inferred from these 386 purged alignments using protein maximum likelihood as implemented in PROTML using the JTT-F matrix with the neighbour-joining tree of maximum-likelihood distances as the starting topology. Of course, not all of the proteins were present in all genomes. Because the alignments with only three sequences had been thrown out (Fig. 1), the smallest number of proteins per alignment was four (there were 18 such trees) and the maximum number of proteins per alignment was 19. The average number of proteins per alignment was 10.6; the average number of sites per alignment after excluding gaps was 186. In total, over 70 000 amino acid positions were considered, but many of them admittedly belong to poorly conserved proteins (see below).

What assumptions are involved here?

The simplest criterion for inferring a cyanobacterial origin of a nuclear-encoded *Arabidopsis* protein would be a common branch for the *Synechocystis* and *Arabidopsis* proteins in a phylogenetic tree, regardless of how the rest of the tree hatches out, as outlined above and in Fig. 2a. But expecting to find this branch for a genuinely cyanobacterial gene in the *Arabidopsis* genome entails quite a few assumptions that are usually made implicitly. It is worthwhile spelling them out.

The first of these assumptions is that methods of phylogenetic reconstruction will in fact reconstruct the common branch for *Synechocystis* and *Arabidopsis* in the case of real data for a nuclear gene that genuinely was acquired from the cyanobacterial antecedent of plastids. Phylogenetic methods have been extensively tested with computer-generated molecular data[9]. But in the case of real data from genuine genomes, it is not always easy to tell whether phylogenetics works reliably or not[10], thereby adding an element of uncertainty to the interpretation of trees in general.

The second assumption is that genes donated from the plastid genome to the nucleus have preserved sufficient sequence conservation to even be detectable with standard database searching programs, so that one can think about making a tree in the first place. Prominent examples of such poorly conserved, but clearly cyanobacterial, proteins in higher plants include the light harvesting complex (LHC) proteins of the thylakoid membrane[11] and TOC75, an outer-envelope component of the chloroplast protein-import machinery[12], both of which share only residual amino acid identity over very short regions with their cyanobacterial counterparts. Such proteins will go largely undetected in our analysis.

The third – and arguably most severe – assumption is that no lateral gene transfer has occurred between the free-living descendants of the cyanobacterial ancestor of plastids and other free-living prokaryotes in the ~1–2 billion years that have passed since plastids arose. This is indeed a very severe assumption because lateral transfer is well known to occur at appreciable rates today[13,14], therefore we should assume it also to have occurred in the distant past. As briefly outlined below, lateral transfer between free-living prokaryotes subsequent to the origins of organelles adds an ominous and presently difficult-to-quantify shroud of uncertainty when it comes to identifying the origin of eukaryotic genes[15]. If neglected or brushed aside[16], it can easily lead one astray.

The fourth assumption is that the product of a gene transferred from the plastid to the nucleus does not alter its function in the cell in a manner that accelerates its rate of sequence evolution so that it is no longer at all similar to the prokaryotic progenitor. Altered function

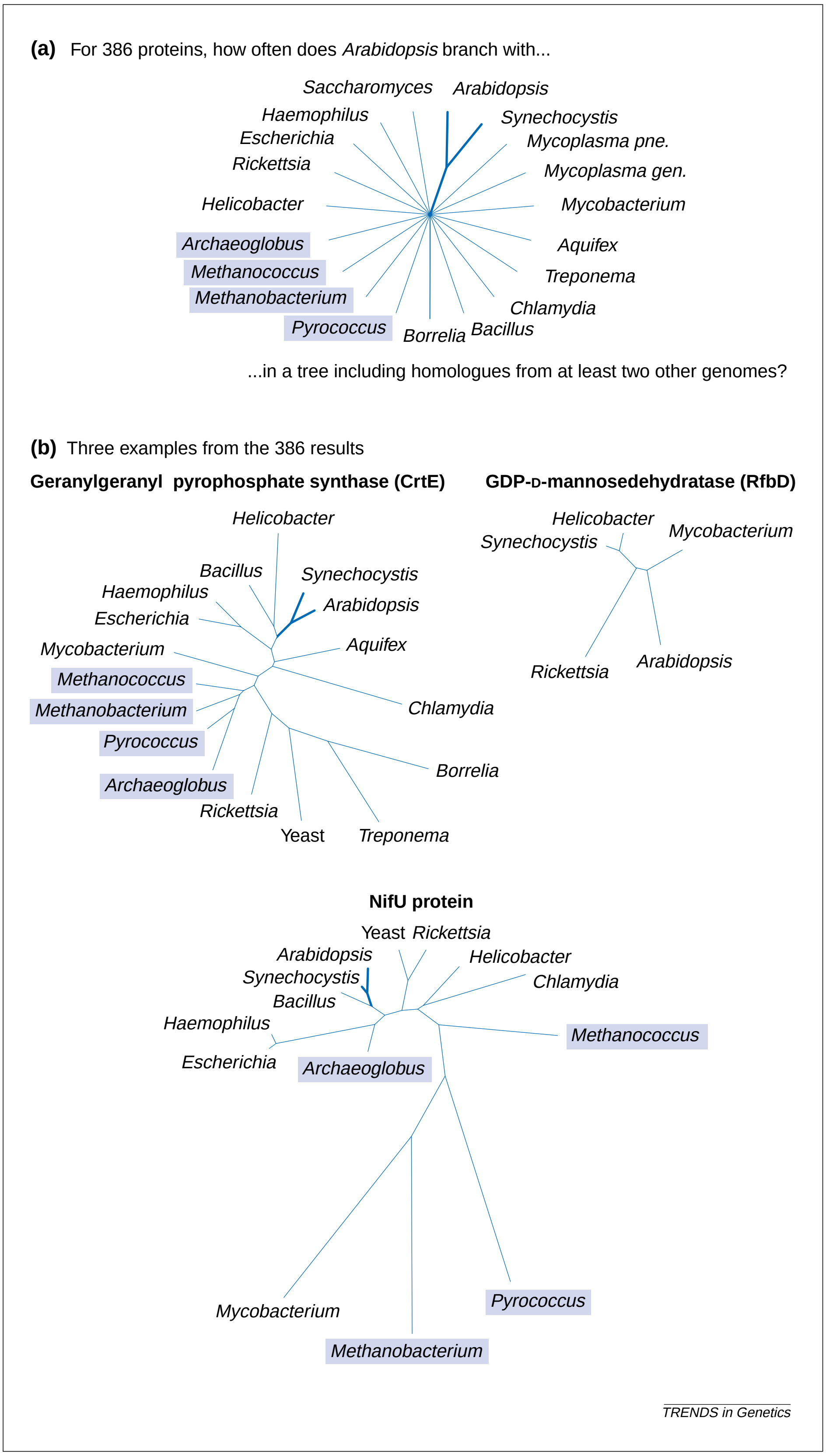


Fig. 2. In search of cyanobacterial proteins in the *Arabidopsis* genome. (a) The genomes that we analysed, indicating the criterion that we used for scoring an *Arabidopsis* nuclear gene as being of cyanobacterial origin; that is, a common branch in phylogenetic analysis regardless of how the other sequences branch (grey area at the centre of the tree). This seemingly simple criterion entails a number of assumptions (see text). The common branch for *Arabidopsis* and *Synechocystis* [designated here as (*Ath*,*Syn*), using standard phylogenetic shorthand] is highlighted, the sequences from archaebacteria are underlined. (b) Three examples of trees that were found, two that indicate a cyanobacterial origin for the *Arabidopsis* protein (CrtE and NifU) and one (RfbD) that does not.

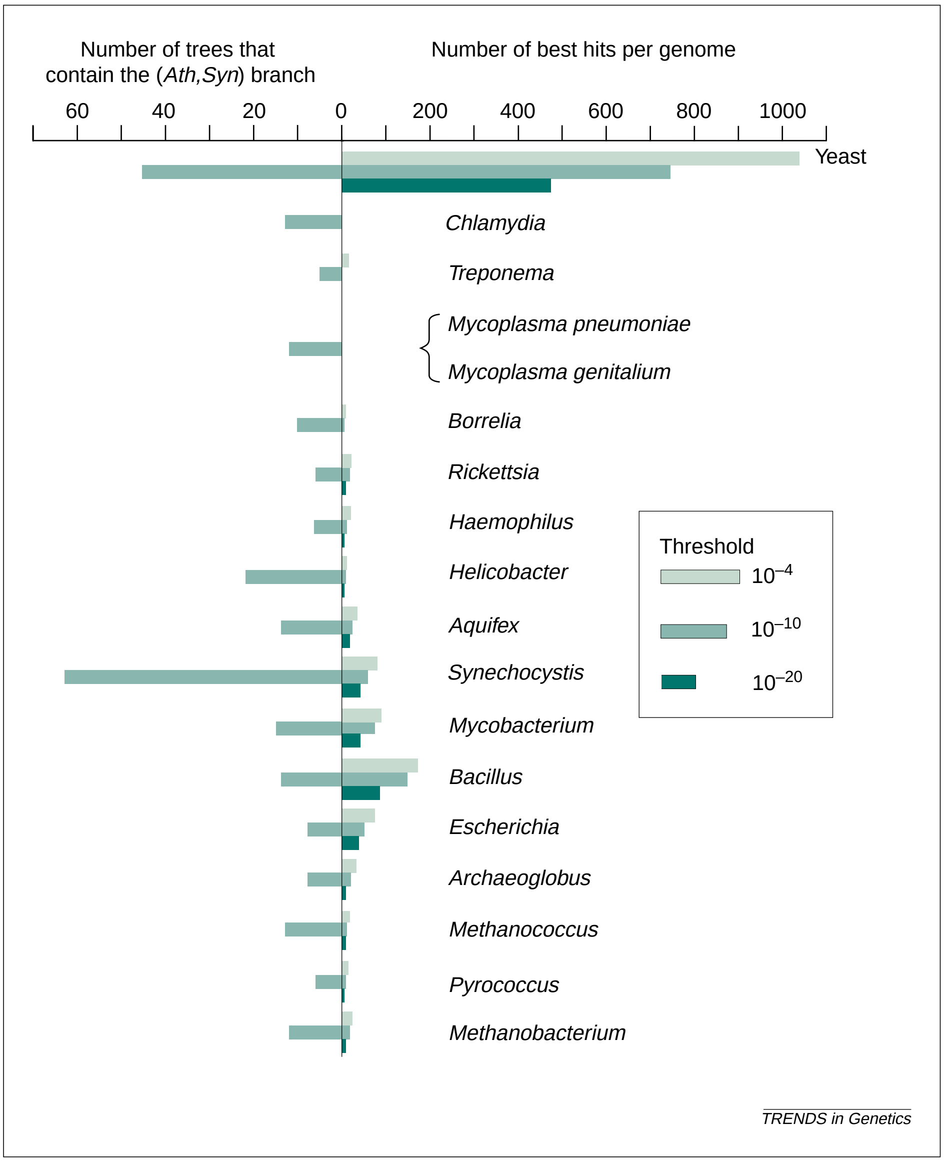


Fig. 3. BLASTs and branches. Bar graphs indicating the number of times that the respective *Arabidopsis* query (3961 proteins) gave the best hit to a protein from the respective genome in BLAST searches (right) and the number of times that a homologue from the given genome branched with the *Arabidopsis* homologue among 386 trees studied (left). BLAST results are shown for three different thresholds. In 110 trees, the *Arabidopsis* homologue did not share a unique common branch with an individual sequence, rather it branched more deeply in the tree. The two *Mycoplasma* species were counted as one genome.

can also result in the accumulation by the nuclear intruder of numerous substitutions that its prokaryotic homologues would not tolerate, leading to what phylogeneticists call a long branch, whereby the correct phylogenetic placement of such long branch sequences in trees is a notoriously difficult problem[17].

In the case of real data from real genomes, the foregoing assumptions are not very realistic. But they *are* the basis of the face-value expectation that plant genes of cyanobacterial origin should branch specifically with the *Synechocystis* homologue in a phylogenetic tree.

A brief summary of 386 protein phylogenies

Given these considerations, what did we find? Sixty-three of 386 alignments investigated yielded a topology in which the *Synechocystis* and *Arabidopsis* homologues shared a unique common branch [designated here as (*Ath*,*Syn*), using standard phylogenetic shorthand]. Two such examples are shown in Fig. 2b, CrtE and NifU, along with one example, RfbD, that did not contain the (*Ath*,*Syn*) branch. The remaining 323 trees (386 minus 63) did not contain the (*Ath*,*Syn*) branch.

Are the 323 proteins whose trees did not contain the (*Ath*,*Syn*) branch direct nuclear inheritances from the host that acquired plastids? If so, they should branch with the yeast homologue, the only other eukaryote in our sample. A glance at the summary of the BLAST results showing the numbers of times that the best hit with the *Arabidopsis* query was found in a particular genome at three expectation thresholds reveals that yeast indeed produced far more ‘best hits’ than any prokaryote sampled (Fig. 3). This would suggest at first sight that the majority of *Arabidopsis* genes sampled were contributed by the eukaryotic host, similar to previous BLAST results reported[5]. But what do the phylogenies say about the 356 out of 386 (92%) cases in which a yeast *and* a *Synechocystis* homologue were present in alignments that we investigated using PROTML? Using the same criterion for gene origin in the case of yeast proteins as we used in the case of *Synechocystis*, that is, a common branch for the yeast and *Arabidopsis* homologues, only 45 of the 386 alignments yielded a unique common branch for *Arabidopsis* and yeast (Fig. 3). Yeast was present in 50 of the 63 trees that indicated a cyanobacterial origin of the *Arabidopsis* gene and it should also be noted that in 110 trees the *Arabidopsis* protein did not branch with any one specific homologue, but rather branched more deeply in the tree.

Among the 386 trees, there were 30 for which no yeast homologue was present. These are genes that *Arabidopsis, Synechocystis* and at least one other prokaryote possess, but yeast does not, making them good candidates for genes of cyanobacterial origin. Indeed, 16 of these gave the (*Ath*,*Syn*) branch.

Returning to the 63 proteins whose trees produced the (*Ath*,*Syn*) branch, we found that these are spread out reasonably uniformly among the various functional classes, as shown in Table 1, nine being found in the class of regulatory functions. So it seems that even some of the regulatory functions encoded in the *Arabidopsis* nucleus were acquired from plastids. Indeed, a few regulatory functions such as bacterial two-component systems, which regulate gene expression, are still encoded in plastid DNA[18].

Adding to those 63 proteins the 11 for which *Synechocystis* had the only detectable homologue to *Arabidopsis* gives 74 proteins among the 3961 studied that seem to come from cyanobacteria (Fig. 1) by the criterion of the (*Ath*,*Syn*) branch. But, as seen in Fig. 3, the

Arabidopsis protein branched on average ten times with homologues from each prokaryotic genome, probably by chance (see below). Subtracting ten from the *Synechocystis* total leaves 64 genes of cyanobacterial origin among the 3961 proteins in the sample.

For the remaining 323 (386 minus 63) trees, we performed a Kishino–Hasegawa test[19] by comparing the log likelihood of the PROTML topology obtained in our first round of tree-building (see Fig. 1) with the log likelihood of a topology in which the *Arabidopsis* homologue was manually forced onto the *Synechocystis* branch. This is outlined in Fig. 4a, where we take the original PROTML tree (Tree1) for a given Protein X, trim off the *Arabidopsis* branch and graft it back onto the tree but now on the *Synechocystis* branch, yielding Tree2, which contains (*Ath*,*Syn*). By calculating the log likelihoods of the two trees and by comparing them in a likelihood ratio test, one obtains a criterion to see whether the trees are significantly different at the 95% confidence level[19].

This test thus provides an estimate of whether the data in the alignment for Protein X would exclude a cyanobacterial origin of the *Arabidopsis* gene for under the maximum-likelihood model, even though the (*Ath*,*Syn*) was not present in the tree initially found by PROTML (Tree1). For 299 of the 323 trees so tested, the topology in which *Arabidopsis* was forced to branch with the *Synechocystis* homologue was not significantly different at $P = 0.95$ from the original PROTML topology.

In other words, of the 386 proteins where a cyanobacterial origin of the *Arabidopsis* protein might be easily detectable, about 64 (1.6% of 3961) suggested a cyanobacterial origin of the *Arabidopsis* gene by virtue of the (*Ath*,*Syn*) branch. But at the same time, fully 299 others do not exclude a cyanobacterial origin by the same criterion at $P = 0.95$, leaving the possibility that as many as 363 (9.2%) might stem from cyanobacteria, because the data do not discriminate.

This is a large margin of uncertainty. Part of the uncertainty might be due to the strength of the Kishino–Hasegawa test[19]; although it is a convenient and well-accepted procedure in phylogenetic studies, its statistical basis has been questioned[20]. But part of this uncertainty might also be attributable to the data, most notably the poorly conserved proteins present in these data. To check that, we plotted the distribution of numbers of trees per category of protein variability to see whether the trees that yield the (*Ath*,*Syn*) branch are among the more highly or more poorly conserved proteins (Fig. 4b). As a rough-and-ready measure of protein variability, we took the total number of substitutions per site in the tree divided by the number of sequences in the tree. For example, if Protein X is twice as variable as Protein Y, it will evolve twice as fast and accumulate twice as many substitutions per site over time. Consequently, the total length of branches in its tree will then be about twice as long as in the tree of Protein Y. So if we compare the total length of the tree for X and Y, we have a rough measure of protein variability – but only if X and Y have the same number of sequences per tree. Because the 386 trees that we are comparing do not always have the same number of sequences in them, ranging from 4 to 19 (see above), we divide by the number of sequences to make these estimates of protein variability more comparable. As seen in Fig. 4b, the vast majority of trees in which the (*Ath*,*Syn*) branch was found belong to the less variable, more conservatively evolving proteins.

Table 1. Functional categories of *Arabidopsis* nuclear genes indicating cyanobacterial origin

Functional category[a]	Number of proteins in the corresponding functional class			
	among the 410 proteins studied here	among trees with (*Ath*,*Syn*) branch	in the *Synechocystis* genome	encoded in any cpDNA[b]
Amino acid biosynthesis	13	1	84	9
Biosynthesis of cofactors, prosthetic groups, carriers	23	4	108	15
Cell envelope	2	1	63	1
Cellular processes	24	1	69	9
Central intermediary metabolism	6	2	31	0
Energy metabolism	15	1	86	3
Fatty acid, phospholipid, sterol metabolism	8	2	36	5
Photosynthesis and respiration	14	6	138	70
Purines, pyrimidines, nucleosides, nucleotides	7	2	39	1
Regulatory functions	103	9	136	3
DNA replication, restriction, modification, recombination, repair	7	0	50	1
Transcription	8	3	23	5
Translation	37	8	155	57
Transport and binding proteins	23	4	158	4
Other categories	51	6	224	7
Hypothetical	69	13	1426	65

[a]Categories refer to those used by Kaneko *et al.*[2]

[b]Data taken from Refs 1,3

This is more evident in Fig. 4c, where the proportion of trees in which the (*Ath*,*Syn*) branch was found is plotted for the categories of protein variability. For clarity, variability intervals of 0.4 on the abscissa in Fig. 4c were used, as compared to intervals of 0.1 in Fig. 4b. In the same figure, the proportion of trees that do not exclude the (*Ath*,*Syn*) branch at $P = 0.95$ using the Kishino–Hasegawa test is also plotted for the same intervals. These results indicate rather clearly that, as protein variability increases, there is a simultaneous decrease in both the chance of recovering a (*Ath*,*Syn*) branch and the ability of the data to discriminate between a cyanobacterial origin or not using the Kishino–Hasegawa test.

These findings indicate that the limiting factor in obtaining a reliable estimate of the number of cyanobacterial genes in higher plants (using *Arabidopsis* as a model) is the degree of protein sequence conservation in such genes that were donated to the nucleus. This is noteworthy, because the issue of how many eukaryotic genes ultimately come from organelles is crucial to understanding how eukaryotic genomes arose[21]. For example, there have been recent claims that as few as 47 of the

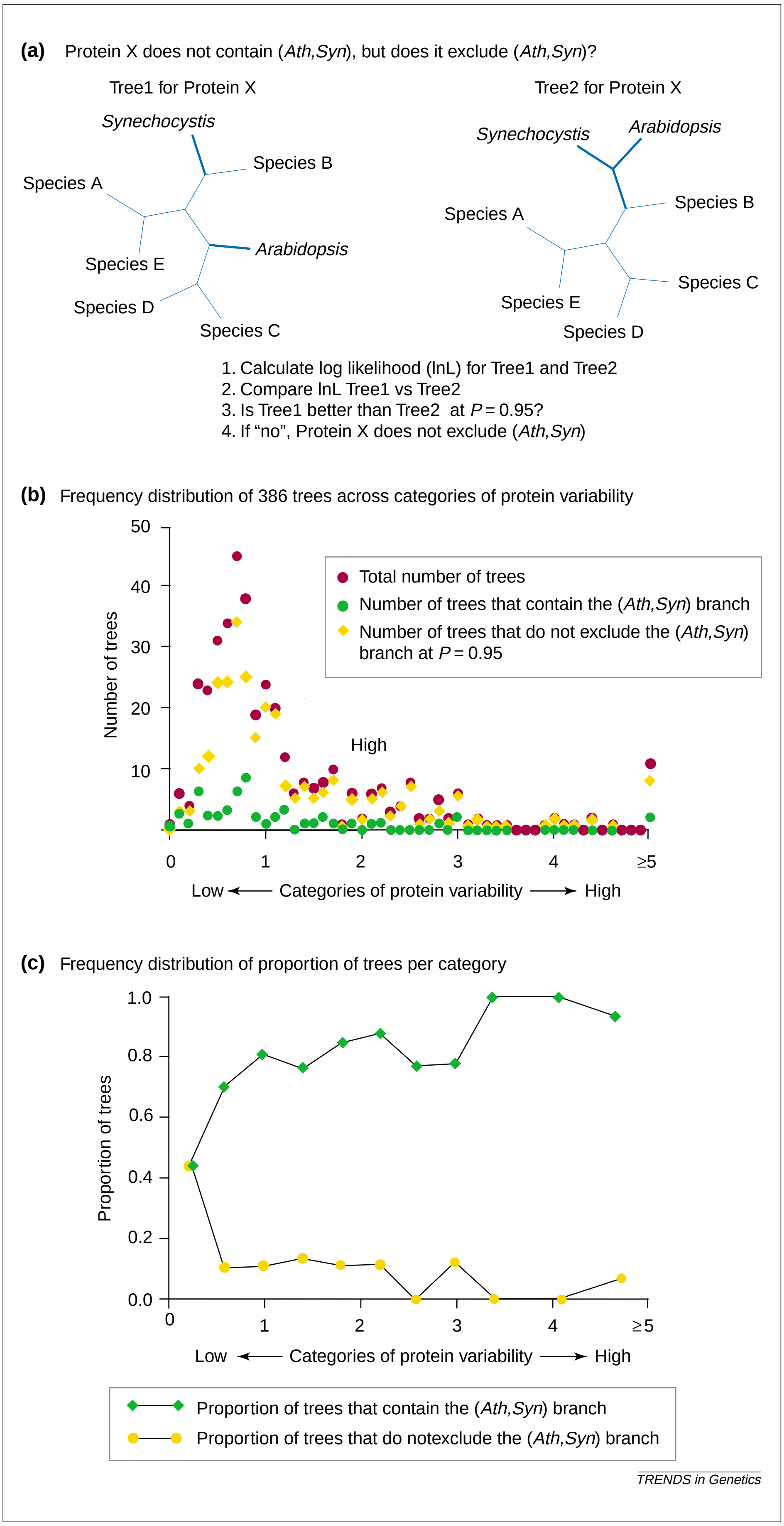


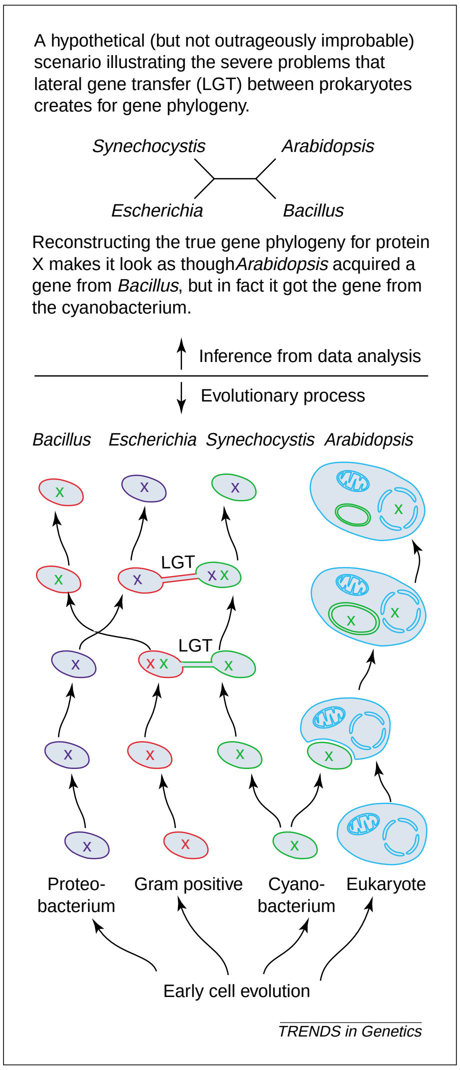
Fig. 4. Sifting through equivocal trees. (a) The Kishinio–Hasegawa test[19] as implemented here for an individual protein (Protein X). Tree1 is the tree originally obtained in phylogenetic analysis, Tree2 is the tree created by moving the *Arabidopsis* branch for comparison with Tree1 (see also Fig. 1). (b) Frequency distribution of 386 trees among categories of protein variability. For each category, the total number of trees, the number of trees that contain the (*Ath,Syn*) branch in Tree1, and the number of trees that do not exclude the (*Ath,Syn*) branch by virtue of the Kishino–Hasegawa test are indicated. Intervals of protein variability of 0.1 on the abscissa were used. Protein variability was measured by taking the total length of Tree1 as measured in amino acid substitutions, dividing by the number of sites in the protein and dividing again by the number of sequences in the tree (see text). (c) Frequency distribution of the fraction of respective trees within each 0.4 interval of protein variability that contain the (*Ath,Syn*) branch in Tree1 (lower curve) and that do not exclude the (*Ath,Syn*) branch when Tree1 and Tree2 are compared (upper curve). Data, alignments, results and Kishino–Hasegawa tests are available at ftp://ftp.134.99.128.42/pub/local/martin

>400 nuclear-encoded proteins of the yeast mitochondrion are acquisitions from α-proteobacteria[16]. But this could be a severe underestimate because there are another 180 nuclear-encoded mitochondrial proteins in yeast with good eubacterial homologues that do not currently reveal an α-proteobacterial origin, hence belonging to what we would call the 'equivocal' class. So for both chloroplasts and mitochondria, sifting through the proteins of the equivocal class (Fig. 1), which our findings show to be predominantly nested among the less conserved proteins (Fig. 4c), will be important for obtaining reliable estimates of the numbers of genes that organelles have donated to the nucleus over time.

Branches, trees and horizontal gene transfer

A cyanobacterial branching for some plant nuclear genes makes good biological sense because it is well known that plastids descend from cyanobacteria that took up permanent residence within their host and transferred many genes to the nucleus[1,3]. So when we see an *Arabidopsis* protein branching with a cyanobacterial homologue in a tree, we infer a cyanobacterial origin of the plant nuclear gene – no problem. But, applying exactly the same logic, for example, to the *Bacillus* and *Mycobacterium* genes that branch with *Arabidopsis* (Fig. 3), we would reach the conclusion that either some plant organelles descend from these Gram-positive eubacteria or that plants acquired a number of their genes through lateral transfer from these Gram-positive eubacteria. In fact, using the face value criterion of common branching, every prokaryote sampled here would appear to have contributed genes to *Arabidopsis* (Fig. 3). And as larger numbers of prokaryotic genomes become available for comparison, countless numbers of prokaryotes would be implicated as having independently donated genes laterally to *Arabidopsis* (or other eukaryotes). This somehow seems unlikely to be true.

In this regard, it should be kept in mind that horizontal gene transfer between free-living eubacteria subsequent to the origins of organelles can indistinguishably mimic outright lateral gene transfer to eukaryotes[15]. As sketched in Fig. 5, if a free-living descendant of the cyanobacterium that gave rise to plastids transferred a gene X to a Gram-positive bacterium (say, a forebear of *Bacillus*) and subsequently had its own gene X replaced by an intruding gene X from proteobacteria (say, *Escherichia coli*), the resulting tree of contemporary genes, if phylogeny is working properly and if the plastid gene was transferred to the nucleus, would have *Arabidopsis* branching with *Bacillus* and *Synechocystis* branching with *E. coli*. This kind of odd branching is very common in phylogenetic trees involving large samples of prokaryotic sequences (see Ref. 22 for several examples and a discussion). In fact, a case similar to this hypothetical example involving a sugar-synthesizing gene in plants and *Chlamydia* was recently reported and taken as evidence for a lateral transfer from *Chlamydia* to plants[23], although the bacterial part of that tree carried the sure signature of lateral gene transfer between bacteria[23], namely the interweaving of species from disparate bacterial groups. Given the prevalence of horizontal gene transfer among eubacteria today[13,14], discriminating in such cases is anything but simple.


A hypothetical (but not outrageously improbable) scenario illustrating the severe problems that lateral gene transfer (LGT) between prokaryotes creates for gene phylogeny.

Reconstructing the true gene phylogeny for protein X makes it look as though*Arabidopsis* acquired a gene from *Bacillus*, but in fact it got the gene from the cyanobacterium.

Fig. 5. Lateral gene transfer – what a problem for phylogenetics! How lateral gene transfer between prokaryotes subsequent to the origins of organelles can lead to erroneous inferences of eukaryotic gene origins (for a discussion see Refs 13,15). In the lower panel, a case of lateral gene transfer (LGT) is depicted as described in the text. The mechanism of LGT sketched here is intended to mean conjugation, but many mechanisms of lateral gene transfer are known[14,15] and for the purposes of the figure, the mechanism is irrelevant. In the upper panel, the tree that would be constructed from those sequences is shown – although the plant obtained its gene from a cyanobacterium, LGT makes it look as though it came from *Bacillus*. As outlined in the text, there is a fine line that separates inferences drawn from phylogenetic data analysis and the evolutionary process itself. Pinning down the role of lateral gene transfer is a very tough problem.

And the standard for comparison for bacterial phylogeny, ribosomal RNA, might not be correct in all cases either. For example, a recent phylogenetic analysis of 9910 amino acids per genome across 39 protein-coding genes for 18 genomes suggested that *Synechocystis* might be more closely related to *Bacillus* and *Mycobacterium* than rRNA phylogeny indicates[24]. The general problem of lateral gene transfer poses a substantial challenge to genome phylogenetics and there is no simple solution in sight. For this and other reasons, it is easier to answer the question of whether a given eukaryotic gene comes from archaebacteria or eubacteria[25,26] than it is to determine exactly which lineage of either it comes from. Despite the problems that it creates, lateral gene transfer can also be useful for phylogenetics. For example, barring loss, once a eubacterial gene has been acquired by a eukaryote, it should be vertically inherited in all descendant lineages. So lateral gene transfer should be of help when it comes to resolving the contours of ancient eukaryotic phylogeny[27], and this, in turn, will help to determine the number of times that independent eubacterial donors made major genetic contributions to eukaryotes. At the minimum, that number will be twice – the origins of mitochondria and plastids[28].

Notwithstanding these vagaries, and coming back to our initial question, projecting the fraction of cyanobacterial genes estimated here onto the *Arabidopsis* genome would give an estimate of the total.

Our assumptions are (1) that the *Arabidopsis* proteins of genuinely cyanobacterial origin are distributed uniformly across the complete spectrum of sequence conservation; (2) that *Arabidopsis* encodes a total of 25 000 proteins[5]; (3) that the evolution of protein families in *Arabidopsis* is equally likely to occur for genes that were acquired from cyanobacteria as for those that were not; (4) that our sample of 3961 proteins is representative for the genome; (5) that *Synechocystis* is a realistic model for the ancestral plastid; and, importantly, (6) that no lateral gene transfer between free-living eubacteria has occurred since the origins of plastids that might influence our inference. Given these, the present analyses suggest that the *Arabidopsis* genome contains between ~400 (1.6%) and ~2200 (9.2%) genes of cyanobacterial origin by the criterion of a unique common branch for the *Arabidopsis* and *Synechocystis* gene. When more data from *Arabidopsis*, other plants, and larger cyanobacterial genomes – such as that of *Nostoc punctiformae* (http://genome.ornl.gov/microbial/npun), with over 7000 proteins – are analysed, we predict that estimates for the number of genes that plants acquired from the cyanobacterial ancestors of plastids will increase.

Acknowledgements

We thank H-W. Mewes for making data available, and K. Henze, M. Hoffmeister and several students for critical comments on the paper. T.R. is currently at Epigenomics AG, Kastanienallee 24, 10435 Berlin, Germany.

References

1 Martin, W. and Herrmann, R.G. (1998) Gene transfer from organelles to the nucleus: How much, what happens and why? *Plant Physiol.* 118, 9–17

2 Kaneko, T. *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3, 109–136

3 Martin, W. *et al.* (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393, 162–165

4 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

5 Abdallah, F. *et al.* (2000) A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci.* 5, 141–142

6 Wolf, Y.I. *et al.* (1999) Rickettsiae and Chlamydiae – evidence of horizontal gene transfer and gene exchange. *Trends Genet.* 15, 173–175

7 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

8 Adachi, J. and Hasegawa, M. (1996) *Computer Science Monographs, No. 28. MOLPHY Version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood*, Institute of Statistical Mathematics, Tokyo

9 Nei, M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* 30, 371–403

10 Lockhart, P.J. *et al.* (1999) Spectral analysis, systematic bias, and the evolution of chloroplasts. *Mol. Biol. Evol.* 16, 573–576

11 Montané, M.H. and Kloppstech, K. (2000) The family of light-harvesting related proteins: was the harvesting of light their primary function? *Gene* 258, 1–8

12 Bölter, B. *et al.* (1998) Origin of a chloroplast protein importer. *Proc. Natl. Acad. Sci. U. S. A.* 95, 15831–15836

13 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2128

14 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304

15 Martin, W. (1999) Mosaic bacterial chromosomes – a challenge en route to a tree of genomes. *BioEssays* 21, 99–104

16 Kurland, C.G. and Andersson, S.G.E. (2000) Origin and evolution of the mitochondrial proteome. *Microbiol. Mol. Biol. Rev.* 64, 786–820

17 Philippe, H. and Germot, A. (2000) Phylogeny of eukaryotes based on ribosomal RNA: Long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* 17, 830–834

18 Race, H.L. *et al.* (1999) Why have organelles retained genomes? *Trends Genet.* 15, 364–370

19 Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order in Hominoidea. *J. Mol. Evol.* 29, 170–179

20 Nei, M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* 30, 371–403

21 Doolittle, W.F. (1998) A paradigm gets shifty. *Nature* 392, 15–16

22 Brown, J.R. and Doolittle, W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* 61, 456–502

23 Royo, J. *et al.* CMP-KDO synthetase: a plant gene borrowed from Gram-negative eubacteria. *Trends Genet.* 16, 432–433

24 Hansmann, S. *et al.* (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: Influence of excluding poorly alignable sites from analysis. *Int. J. Syst. Evol. Microbiol.* 50, 1655–1663

25 Ribiero, S. and Golding, G.B. (1998) The mosaic nature of the eukaryotic nucleus. *Mol. Biol. Evol.* 15, 779–788

26 Rivera, M.C. *et al.* (1998) Genomic evidence for two functionally distinct gene classes *Proc. Natl. Acad. Sci. U. S. A.* 95, 6239–6244

27 Baldauf, S.L. *et al.* (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290, 972–977

28 Martin, W. and Müller, M. (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392, 37–41

T. Rujan
W. Martin*
Institute of Botany III, Universität Düsseldorf, Universitätsstr. 1, D-40225 Düsseldorf, Germany.
*e-mail: w.martin@uni-duesseldorf.de

Were protein internal repeats formed by ‘bricolage’?

Giovanni Lavorgna, La’szlo’ Patthy and Edoardo Boncinelli

Is evolution an engineer, or is it a tinkerer – a ‘bricoleur’ – building up complex molecules in organisms by increasing and adapting the materials at hand? An analysis of completely sequenced genomes suggests the latter, showing that increasing repetition of modules within the proteins encoded by these genomes is correlated with increasing complexity of the organism.

Evolution has brought about the formation of organisms of increasing complexity. This process involved mechanisms, such as exon-shuffling[1] and gene duplication[2], that increased intermolecular duplications of the more sophisticated proteomes. For example, gene duplication contributed to the origin and evolution of vertebrates, which appear to possess several copies of an ancestral set of genes[3]. A single gene in flies usually has three or four paralogous genes in mammals, and this spare genetic capacity has permitted new possibilities, allowing the acquisition of new biochemical functions and expression capabilities[4].

More than two decades ago, when only a handful of eukaryotic genes were cloned, Francois Jacob had already envisioned some of these basic evolutionary mechanisms[5]. In fact, he argued that evolution could work as a tinkerer, rather than an engineer, implying that evolutionary processes construct things with the materials at hand and the outcome bears the constraints imposed by those materials[6]. Translated into molecular terms, the raw materials are the existing set of genes, which can be, in part or entirely, elaborated again and redeployed to a new function during evolution. Extending to Jacob’s view of ‘recyclement’ of biological material, we investigated systematically the possibility that, besides the increase of *inter*-molecular duplications, an increase of *intra*-molecular duplications accompanied the evolution of proteins.

We decided to look for repeated protein modules, as opposed to short, low-complexity sequence repeats (i.e. runs of Qs, STSTSTSTS, etc) because, in several instances, modules of proteins are used to build the function of many multidomain proteins. As a result, we found, with a few exceptions, that:

(1) There is a correlation between the complexity of functions controlled by the proteome of a given organism and its degree of internal repetitiveness.

(2) The above correlation is often observed both for interdomain comparisons (e.g. archaeal proteins have, on average, more internal repeats than bacterial ones) and intradomain comparisons (e.g. human proteins have more internal repeats than those belonging to *Drosophila melanogaster*).

(3) We also detected a decrease in the number of internal repeats following ‘reductive’ evolution, in which the biological complexity of an organism is

http://tig.trends.com 0168–9525/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02207-1