Research

On the species of origin: diagnosing the source of symbiotic transcripts

Peter T Hraber* and Jennifer W Weller†

Addresses: *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA. †Virginia Bioinformatics Institute, 1750 Kraft Drive, Suite 400, Blacksburg, VA 24061, USA.

Correspondence: Peter T Hraber. E-mail: pth@santafe.edu

Published: 23 August 2001

Genome ***Biology*** 2001, **2(9)**:research0037.1–0037.14

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/9/research/0037

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 11 June 2001
Revised: 11 July 2001
Accepted: 25 July 2001

Abstract

Background: Most organisms have developed ways to recognize and interact with other species. Symbiotic interactions range from pathogenic to mutualistic. Some molecular mechanisms of interspecific interaction are well understood, but many remain to be discovered. Expressed sequence tags (ESTs) from cultures of interacting symbionts can help identify transcripts that regulate symbiosis, but present a unique challenge for functional analysis. Given a sequence expressed in an interaction between two symbionts, the challenge is to determine from which organism the transcript originated. For high-throughput sequencing from interaction cultures, a reliable computational approach is needed. Previous investigations into GC nucleotide content and comparative similarity searching provide provisional solutions, but a comparative lexical analysis, which uses a likelihood-ratio test of hexamer counts, is more powerful.

Results: Validation with genes whose origin and function are known yielded 94% accuracy. Microbial (non-plant) transcripts comprised 75% of a *Phytophthora sojae*-infected soybean (*Glycine max* cv Harasoy) library, contrasted with 15% or less in root tissue libraries of *Medicago truncatula* from axenic, *Phytophthora medicaginis*-infected, mycorrhizal, and rhizobacterial treatments. Mycorrhizal libraries contained about 23% microbial transcripts; an axenic plant library contained a similar proportion of putative microbial transcripts.

Conclusions: Comparative lexical analysis offers numerous advantages over alternative approaches. Many of the transcripts isolated from mixed cultures were of unknown function, suggesting specificity to symbiotic metabolism and therefore candidates likely to be interesting for further functional investigation. Future investigations will determine whether the abundance of non-plant transcripts in a pure plant library indicates procedural artifacts, horizontally transferred genes, or other phenomena.

Background

Access to automated DNA sequencing technology has made possible the rapid generation and analysis of gene transcripts expressed in organisms via expressed sequence tags (ESTs) [1-5]. This information has helped to identify those genes expressed in particular stages of development and in specialized tissues or organs [6-10]. Novel gene products and target leads for therapeutic intervention can also be gleaned rapidly from ESTs [1,2,11]. A more detailed understanding of the molecular interactions between symbionts, whether pathogenic or mutualistic [12], is also possible with this approach [13-17].

For a sequence isolated from interacting symbionts, determining its cellular role (or roles) is complicated by not knowing which species expressed the sequence [18]. We refer to this challenge as 'the problem': given a sequence x expressed in an interaction between species A and B, did x originate from A or B? Various solutions are readily conceived, each with merits and faults. Here, we show that a comparative lexical analysis of word counts (specifically, hexamer frequencies), previously used to detect library contamination in sequencing projects [19], provides a powerful computational basis to infer a transcript's species of origin.

Experimentally, one can attempt to solve the problem by hybridizing a clone (as probe) to genomic DNA (target) from both species and determining to which target the probe hybridizes. This approach can produce very reliable results. However, if a sequence is highly conserved in the two taxa, hybridization stringency conditions can influence the outcome considerably. For high-throughput EST sequence analysis, source verification by hybridization is impractical in terms of time and reagents. As an alternative to *in vitro* hybridization, several computational solutions are possible.

Were the genome sequence of both species completely determined, one could simply use sequence similarity searching [20-22]. However, most plant hosts and their microbial symbionts have little or no genomic sequence data available, which makes this approach very unreliable. Strong similarity to a sequence from one organism does not preclude the possibility that a similar sequence is present in the other species. Conclusions based upon such partial knowledge have been informative, but are potentially misleading [18,23].

Codon usage varies across taxa [24-26]. Exploiting this fact may seem a viable solution to the problem, as it has proven suitable for predicting the presence of introns among exons in genomic DNA. However, it really is not practical, because of the need to know the reading frame for translation of a messenger RNA into an amino acid. EST data are of notoriously unreliable quality, sometimes having a large proportion of ambiguous bases, and sometimes having single base-pair insertions or deletions, which disrupt a reading frame. Word counting is less prone to these sources of error, and uses information intrinsic to biases in codon usage by counting codon pairs as hexamers in a sliding window, whereas codons are read in non-overlapping, tiled windows.

An intuitive approach to the problem that examines sequence composition is to compare the guanine and cytosine (GC) base content of a sequence with other sequences from the species being studied. When two species' genomes have different GC content, this method can be very useful. In a recent investigation, for instance, sequences from the stramenopile plant pathogen *Phytophthora sojae* and its soybean (*Glycine max*) host showed a 20% difference in mean GC content [18]. The origin of a number of sequences could readily be identified this way, but a large proportion could not, because of considerable overlap in the distributions' tails. Counting frequencies of GC is simple word counting, where the word size k is 1/2: only two semi-words, G/C and A/T are counted.

An alternative approach to determining the origin of a sequence is suggested by previous work on analysis of word counts, or k-tuple frequencies, which was intended as a means of evaluating a library for contamination when sequencing from a single model organism [19]. The word-counting method provides distinct advantages over other computational methods. Unlike sequence-similarity searching, it does not require that the full protein-coding content of both genomes be known for reasonable inferences to be made. Further, word counting is sensitive to biases in codon usage and GC content commonly observed when comparing taxa, but does not require knowledge of the reading frame for amino-acid translation. That is, the underlying differences between the two organisms that result in base composition or codon usage biases can also be detected by counting words. Unlike GC analysis, lexical analysis establishes a clear threshold above or below which we can infer the species of origin, and a confidence level for an inference can readily be assigned. Dunning's likelihood-ratio test of word dissimilarities [27] also has the appealing property of being non-parametric, having no assumption of normality for the underlying frequency distribution, which makes it statistically powerful [28]. Dunning [27] demonstrated that unreliable results can be obtained from parametric tests, such as χ^2, particularly in such cases as lexical analysis.

In the experiments detailed below, we first validate the word-counting method on sequences whose origin and function are known, then compare it with ability to diagnose the origin of sequences with distributions of GC content. We examine sequences from pathogenic interactions between species from the genus *Phytophthora* and the plant hosts *G. max* and *Medicago truncatula*, then apply the word-counting approach to sequences from two microbial mutualists in association with *M. truncatula*, the arbuscular mycorrhizal zygomycete *Glomus versiforme*, and the nitrogen-fixing bacterium *Sinorhizobium meliloti*.

Results

Validation sequence accession numbers, gene names, and comparison results appear in Table 1. Incorrect inferences are underlined. The word-counting method was generally quite reliable when tested against sequences of known origin, being wrong in 3 cases out of 50; a phosphate transporter from *G. versiforme* and two *in planta*-induced genes from *Phytophthora infestans* were misidentified as plant sequences. This indicates a failure rate of 6% - all false negatives under the null hypothesis that a transcript originates from the plant host. Performance of the method was not

Table 1

Dissimilarity (*D*) comparison results from 50 validation sequences

Accession	Gene name	mRNA (?)	Length (nucleotides)	*D*(*A*) plants	$D(B_1)$ oomycetes	$D(B_3)$ bacteria
Glomus versiforme						
AJ009628	chitin synthase *Gvchs1*	n	638	2,535.2	2,468.6	2,718.4
AJ009629	chitin synthase *Gvchs2*	n	481	2,203.2	2,050.0	2,286.0
AJ009630	chitin synthase *Gvchs3*	n	4,116	7,205.9	5,235.8	5,985.8
U38650	phosphate transporter	y	1,833	3,937.9	5,702.3	6,514.3
Glycine max						
J01297	actin *SAc3*	n	1,620	3,322.0	4,554.6	5,329.7
K00821	lectin *Le1*	n	2,152	4,124.6	6,558.3	7,928.3
M64267	iron superoxide dismutase	y	1,056	2,773.6	3,761.2	4,269.2
Medicago truncatula						
AF000354	phosphate transporter *MtPT1*	y	1,920	3,800.3	5,630.7	6,654.2
AF000355	phosphate transporter *MtPT2*	y	1,867	3,673.9	5,390.1	6,424.0
AF055921	*Mt4* genomic sequence	n	954	2,631.9	4,004.4	4,539.1
AF106929	cell wall protein *AM1*	y	885	3,433.6	4,200.0	4,774.3
AF106930	translation initiation protein *AM3-1*	y	3,154	4,557.6	5,982.7	7,212.8
AF106931	translation initiation protein *AM3-2*	y	1,384	3,371.1	4,130.0	4,644.4
AJ132891	*ha1* gene, exons 1-22	n	3,620	4,383.2	8,683.6	10,730.7
AJ388847	*MtNo213* superoxide dismutase	y	530	2,110.2	2,219.8	2,367.0
AJ388865	*MtNo233* triosephosphate isomerase	y	563	2,171.6	2,405.6	2,618.6
U16727	peroxidase precursor *rip1*	n	2,603	4,246.1	8,210.0	9,901.9
U38651	sugar transporter	y	1,728	3,619.6	5,128.4	5,976.5
X57732	leghemoglobin *Mtlb1*	n	1,073	3,021.3	5,029.1	5,845.9
X57733	leghemoglobin *Mtlb2*	n	592	2,045.9	3,156.0	3,568.2
X60386	lectin *lec1*	n	1,363	3,228.8	4,935.4	5,605.6
X60387	lectin *lec2*	n	1,192	3,142.8	4,472.6	4,985.9
X82216	*lec3*	n	1,155	2,928.4	4,283.3	4,930.8
X68032	*ENOD12*	n	772	2,780.4	3,679.7	4,096.5
X99466	*ENOD16*	n	1,142	3,124.6	4,535.5	5,156.2
X99467	*ENOD20*	n	1,405	4,003.6	5,294.7	5,966.7
Y10267	glutamine synthetase	y	1,413	3,116.1	4,506.5	5,292.1
Y10373	chitinase	y	1,305	3,369.5	4,090.4	4,703.4
Phytophthora infestans						
AF004951	surface glycoprotein elicitor *inf2A*	y	648	3,428.4	2,421.9	2,589.1
AF004952	surface glycoprotein elicitor *inf2B*	y	701	3,611.7	2,514.5	2,698.6
L23938	*ipiO2*	n	1,556	4,125.2	4,339.5	4,855.5
L23939	*ipiO1*	n	1,826	4,360.5	4,580.9	5,259.0
L24206	*ipiB1*	n	1,726	6,086.7	4,584.3	5,159.3
M59715	actin *actA*	n	1,736	5,137.1	3,637.2	4,420.0
M59716	actin *actB*	n	1,405	4,425.3	3,569.5	4,141.6
M83535	calmodulin *calA*	n	1,358	4,063.0	3,724.9	4,138.1
X64537	*tigA*	n	2,448	6,221.0	4,193.8	5,181.9
P. capsici						
U42304	chitin synthase *chs*	n	449	2,238.8	1,882.7	1,997.5
P. parasitica						
X97205	cellulose-binding-elicitor lectin	y	918	3,819.1	2,876.0	3,208.4
Sinorhizobium meliloti						
AF040724	*nodD*	n	1,776	5,317.9	4,197.8	4,179.4
AF110770	superoxide dismutase *sodA*	n	1,196	4,898.2	3,343.2	2,916.0
M61753	*exoD*	n	858	4,071.8	2,847.2	2,372.9
M68858	nodulation proteins *nodP* and *nodQ*	n	3,476	9,992.3	5,288.0	3,954.7
M96261	phosphate regulators *phoU* and *phoB*	n	1,178	5,332.7	3,359.3	2,866.4
U90221	*syrA*	n	1,102	4,176.2	3,375.8	3,220.8
X01649	*nodA*, nodB, and *nodC*	n	3,373	7,684.1	4,819.5	4,646.1
X03065	regulatory nitrogen fixation *fixD*	n	2,111	5,723.0	4,249.8	4,228.3
X17523	glutamine synthetase II	n	990	4,303.5	2,959.9	2,720.4
Y08500	*putA*	n	3,804	13,623.3	6,376.5	4,212.0
Agrobacterium tumefaciens						
U91632	sugar transporter *gguA* and membrane-spanning permeases *gguB* and *gguC*	n	4,185	11,132.6	5,959.4	4,551.4

Incorrect inferences are underlined.

influenced by whether the isolated source of a sequence was an mRNA or DNA molecule, as indicated by the column labeled 'mRNA?'.

Distributions of GC content are approximately normal in two of three cases studied, those of axenic *P. sojae* cultures (Figure 1). For sequences from infected plant cultures, a bimodal distribution is apparent. Roughly 25% of a total of 927 infected *G. max* sequences contain less than 50% GC; most of these are likely to be plant transcripts [18]. This is a considerably greater number than for axenic *P. sojae* cultures, in which fewer than 5% of mycelia and zoospore isolates contain less than 50% GC.

Several properties of cumulative distribution functions warrant comment, to help explain similar plots from word dissimilarity comparisons (Figures 1b,2a). The median of a distribution occurs where the function reaches a cumulative probability of 0.5. Medians from all three *P. sojae* libraries are similar, varying by less than 4% GC (Figure 1b). Other moments of the distributions are readily apparent; the variance is inversely related to the slope at the median value of the function. A useful property of cumulative distribution functions is that any point on the *y* axis gives the integrated area (cumulative probability) under the curve. We use this property to establish experiment-wide false-positive and false-negative rates (Figure 2a). In this case, $\alpha = 0.088$ and $\beta = 0.032$.

Calibration curves from hexamer dissimilarity tests, shown in Figure 2b as solid black lines for plant and dashed black lines for stramenopile training sequences are approximately normal. The medians differ considerably, with only about 10% percent overlap in the two distributions' tails about the neutral *t*-value of zero. Superimposed are comparison curves from *P. sojae* test sets (Figure 2b), which parallel the GC content curves in Figure 1b but show slightly less variance. Axenic sequences are clearly more like stramenopiles (B_1) than plants (A_1) in hexamer composition, with all but a small percentage having positive *t* values. Plant-like sequences are as abundant in the mixed library as detected by GC content, about 23%. As expected, the two methods agree, having positively correlated values for GC and *t* ($r^2 = 0.852$, $P < 10^{-16}$, $\nu = 2{,}641$).

Looking in more detail at the paired dissimilarity values (Figure 3), we can see which individual sequences are more or less like plant and pathogen. The magnitudes of dissimilarity are also apparent, with longer sequences having larger dissimilarity values. BLASTX similarity searches against the protein sequences in nr, a non-redundant library of proteins [29-31] revealed that none of the 12 plant-like mycelial transcripts significantly resemble known proteins ($E > 10^{-4}$). Among the top ten most plant-like transcripts from the infected *G. max* library, three had no significant matches, four matched putative *Arabidopsis thaliana* proteins, and three matched known *G. max* proteins: cytochrome P450 (accession AF022460, $E < 10^{-34}$), methylglyoxalase (accession P46417, $E < 10^{-34}$), and a ripening related protein (accession AF127110, $E < 10^{-71}$). Thus, the majority of the most plant-like transcripts in the infected soybean library strongly resemble characterized plant sequences. Analysis results from all *P. sojae* and mixed-culture transcripts are available online as additional data files, grouped by the library from which transcripts were sequenced.

Figure 4 shows that calibration curves from comparing plant and microbial symbiont training sets have good separation and minimal overlap (about 10%) in two of three cases, but not for training set B_2, comprised of zygomycetes and chytridiomycetes, which overlaps considerably with plants (Figure 4b). The associated error rates are $\alpha = 0.126$ and $\beta = 0.207$. When comparing between plants and bacteria, the error rates are $\alpha = 0.052$ and $\beta = 0.084$, much lower than when comparing plants (A_2, *Medicago*) with fungi (B_2, zygomycetes and chytridiomycetes). Error rates for comparing stramenopiles and *P. infestans* ESTs with plants are as in Figure 2 ($\alpha = 0.088$, $\beta = 0.032$).

Also shown in Figure 4 are cumulative distributions from comparisons with *M. truncatula* and microbial symbionts. All resemble calibration curves from plant sequences, having similar medians and slightly less variance than the plant calibration curves. Comparison curves show that the great majority of test sequences are more plant-like than otherwise, with 20% or less resembling microbial symbionts more closely than plants. A greater proportion of microbial sequences is present in the *M. truncatula–G. versiforme* interaction library (20%, Figure 4b) than in the *P. medicaginis*-infected *M. truncatula* library (5%, Figure 4a). However, Long's root-hair enriched library (MtRHE) [6] had a greater proportion of putative microbial sequences present (7% and 25%) than any of the libraries isolated from symbiont-associated cultures. The axenic and nodulating root libraries had the smallest portion of putative microbial transcripts (< 2%, Figure 4c), with the axenic library closely resembling nodulating root libraries. The method of preparing a library can affect the proportion of plant and non-plant sequences, as discussed below.

Paired dissimilarity values in Figure 5 show in greater detail which sequences are more or less like plant and symbiont. Sequences from an interaction library and pure plant root cultures appear together for comparison. Considerable variation in the degree of dissimilarity to both training sets is clear, largely due to variation in the length of sequences within test sets. Consistent with the cumulative distributions of $D(A) - D(B)$ in Figure 4, most sequences lie above the identity function, and resemble the plant host more closely than the microbial symbiont. Mycorrhizal test sequences are more difficult to differentiate than sequences from the rhizobacterial or pathogenic associations, as seen by the diminished variation about the identity function in mycorrhizal comparisons (Figure 5b), contrasted with comparisons from

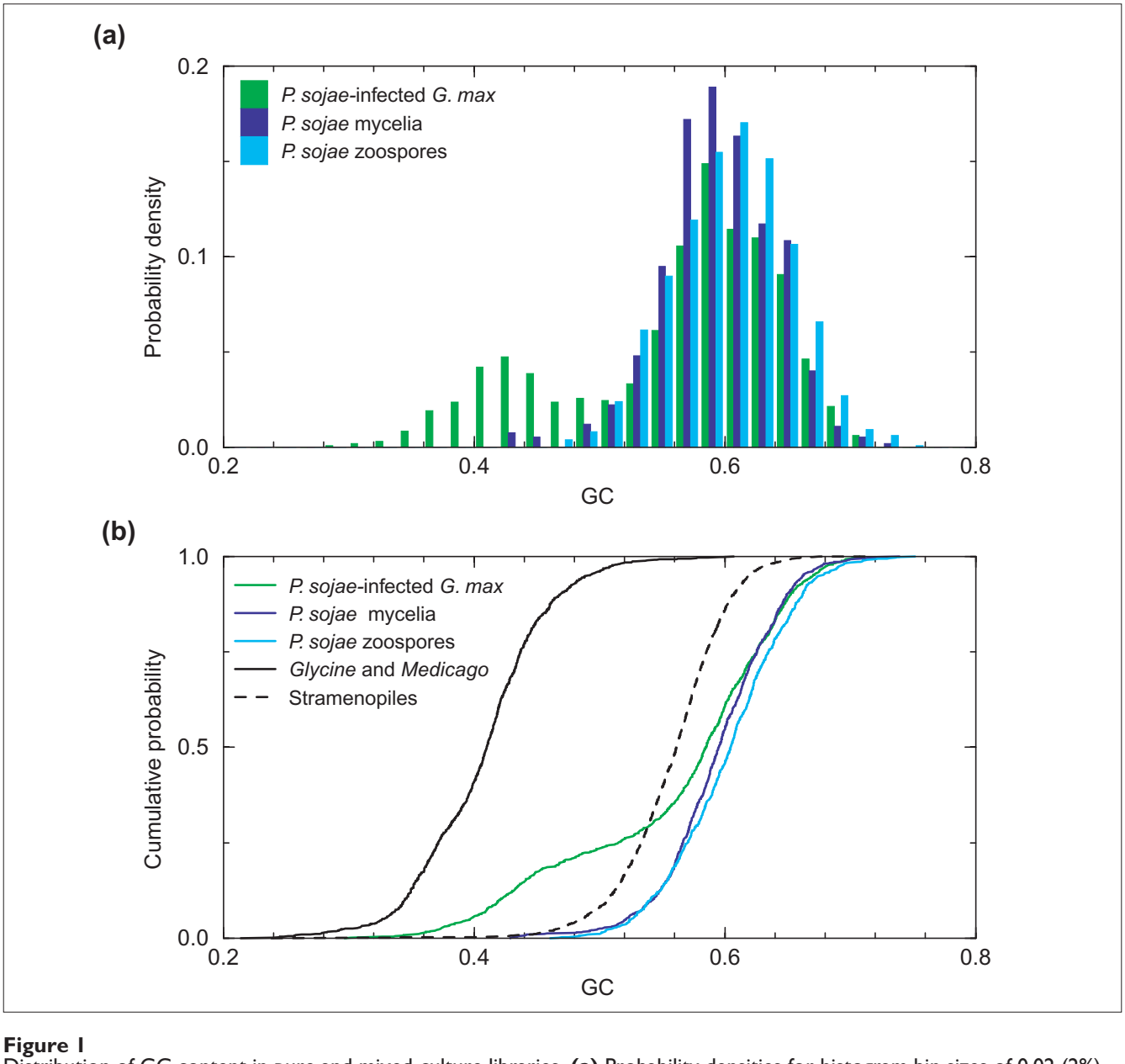

Figure 1
Distribution of GC content in pure and mixed-culture libraries. **(a)** Probability densities for histogram bin sizes of 0.02 (2%) in base content. **(b)** Cumulative probability distribution functions (*cdf*s).

pathogen-infected and nodulating root libraries (Figures 5a and c, respectively). Analysis results from all *M. truncatula* and mixed-culture transcripts are available as additional data files online, grouped by the library from which transcripts were sequenced, and sorted from the least plant-like transcripts to the most plant-like.

Discussion

Clearly, the word-counting approach provides a reliable solution to the problem of source identification with known confidence, and has several significant advantages. The reliability of the method is best justified in terms of the favorable validation test results, and is further corroborated by agreement with an analysis of GC content. In test cases where the correct answer is known *a priori*, results were correct within error rates expected from overlap in training sets. (Recall that $\alpha = 0.088$ for comparisons between plants and stramenopiles, and $\alpha = 0.052$ for comparisons between plants and bacteria.) Unlike GC content, the problem is clearly resolved by word counting with a threshold value of $t = 0$, and with statistical rigor, because false-positive and false-negative rates for a set of comparisons are readily computed from cumulative distributions of dissimilarity between two

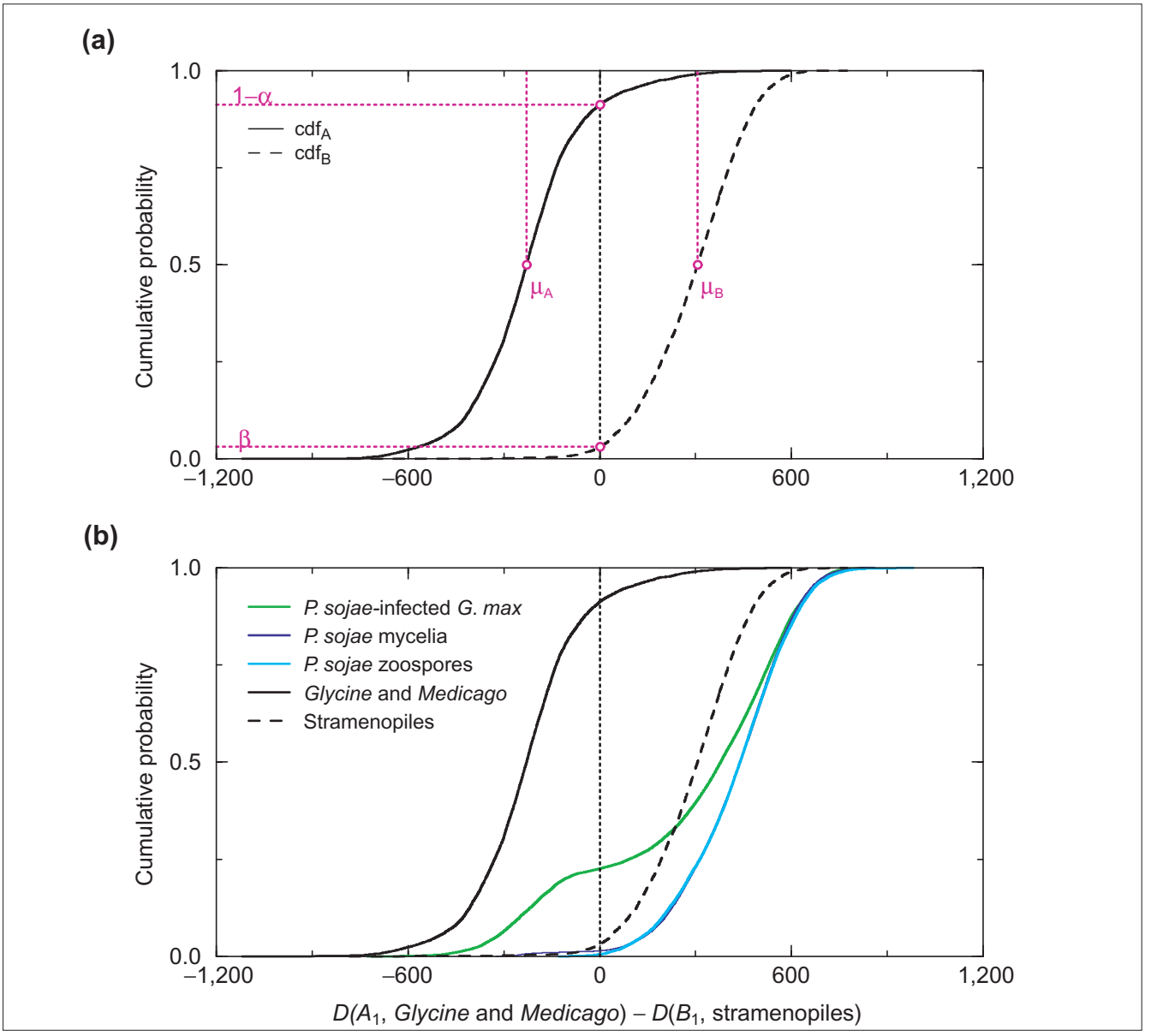

Figure 2
Distribution of hexamer dissimilarity test results from pure and mixed-culture libraries. **(a)** Calculation of statistical parameters from *cdf*s *A* and *B*. Overlap in the upper tail of cdf_A with cdf_B and the lower tail of cdf_B with cdf_A are likely regions for error. We find the false-positive rate α where $1 - cdf_A$ intersects 0 [$cdf_A(0) = 1 - \alpha$], and the false-negative rate β where cdf_B crosses 0. Also shown are the medians (μ) for each distribution, where $cdf(\mu) = 0.5$. **(b)** Calibration curves for plant (A_1, *Glycine* and *Medicago* spp., solid black line) and stramenopile plus *P. infestans* EST (B_1, dashed black line) training sequences. Superimposed distributions of test results show dissimilarity differences for infected *G. max* (green) and axenic *P. sojae* mycelial and zoospore sequences (blue and cyan, respectively).

training sets. Optimal statistical power (minimal false-negative rate) is ensured when using a likelihood-ratio test statistic, as demonstrated by the Pearson-Neyman Theorem [28]. Further, word counting need not be trained only for the species being compared. Rather, it is sufficient that the training set be related to, but not necessarily congeners of, the species from which sequences are being compared. Sequences from several species of the genus *Phytophthora* were correctly distinguished from plant and bacterial sequences, and three genes from *Agrobacterium tumefaciens* were correctly identified as representing a bacterial sequence.

However, several caveats warrant prudence. Transcribed sequences that do not encode proteins, but rather catalytic single-stranded RNAs such as transfer and ribosomal RNAs [32], should be treated independently because they are more

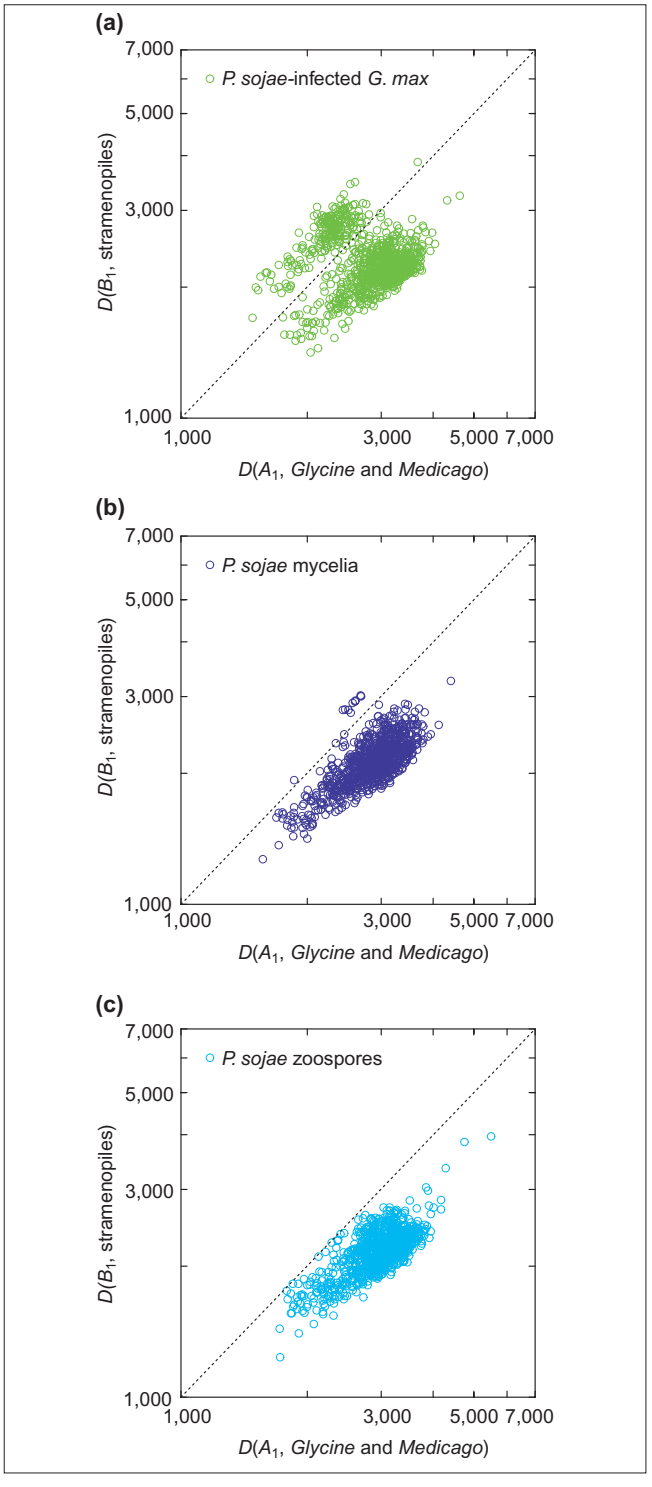

Figure 3
Paired dissimilarity test results from pure and mixed-culture libraries. Each point corresponds to an expressed tag from either **(a)** infected *G. max* or **(b)** axenic *P. sojae* mycelial or **(c)** zoospore sequences, compared with plant (A_1) and stramenopile plus *P. infestans* EST training sequences (B_1). The identity function indicates equal dissimilarity to both training sets, $t = D(A) - D(B) = 0$. Points above the identity function are more plant-like than points below.

highly conserved across taxa than messenger RNAs. Also, filtering or trimming of low-complexity repeat regions, such as poly(A) or poly(T) tracts, is helpful because comparison results can be influenced by the abundance of a single hexamer. Early in our investigations, using one set of training sequences obtained from directionally cloned *P. infestans* cDNAs produced results that were difficult to interpret. It eventually became clear that, as the *P. infestans* sequences were all single-pass reads from the 5′ end of a clone generated with the T3 primer, few sequences complementary to the 3′ end of the mRNA sequence were present in the training set. This meant that the hexamer AAAAAA was common, but the hexamer TTTTTT scarce. Large amounts of the poly(T) hexamer would be expected when sequencing reverse complements of mRNAs obtained from 3′ sequences generated with the T7 primer. Both poly(A) and poly(T) regions were present among plant training sequences. As a result, any sequence that contained a poly(T) tract tended to resemble the plant sequences. Further, because the error rates for an inference depend on the degree to which calibration curves overlap, the best results are obtained where overlap is minimal. Despite these caveats, word counting presents a viable solution to the problem.

The *P. sojae*-infected *G. max* library provides a clear example of contrast in both hexamer composition and GC content, resulting in readily diagnosed origins. Not every case is this simple. For clear separation between the two species to appear, the two must differ in composition and a detectable proportion of transcripts from each species must be present in the library. To be detectable, the proportion of transcripts present from a particular species must be greater than the error rate obtained from calibration curves.

Though these criteria are true for the infected *G. max* library ($t < 0$ for < 25% of 927 transcripts), they do not appear to be true for the *M. truncatula* libraries we analyzed ($t < 0$ for 80-99% of 890-3,017 transcripts). In the *P. medicaginis* interaction library, we might expect the same bimodal distribution as seen with *P. sojae*. However, the two libraries were prepared in different ways. The *P. sojae*-infected library was prepared two days after infection, using a susceptible plant host strain, so as to maximize the number of pathogen transcripts present in the host tissue [18]. Further, *G. max* hypocotyl tissues were infected directly with a zoospore suspension. In contrast, the *P. medicaginis*-infected library was prepared ten days after infection and individual plants varied in their degree of susceptibility (C. Vance, unpublished data). Plants were also inoculated in a different manner: ground mycelia were dissolved in sterile water and incubated, and the resulting inoculum was pipetted onto the soil surface, rather than the plant. These differences in how tissues were cultured prior to library preparation could have produced the disparate abundance of plant transcripts, though both libraries were prepared from plant tissues infected with *Phytophthora*.

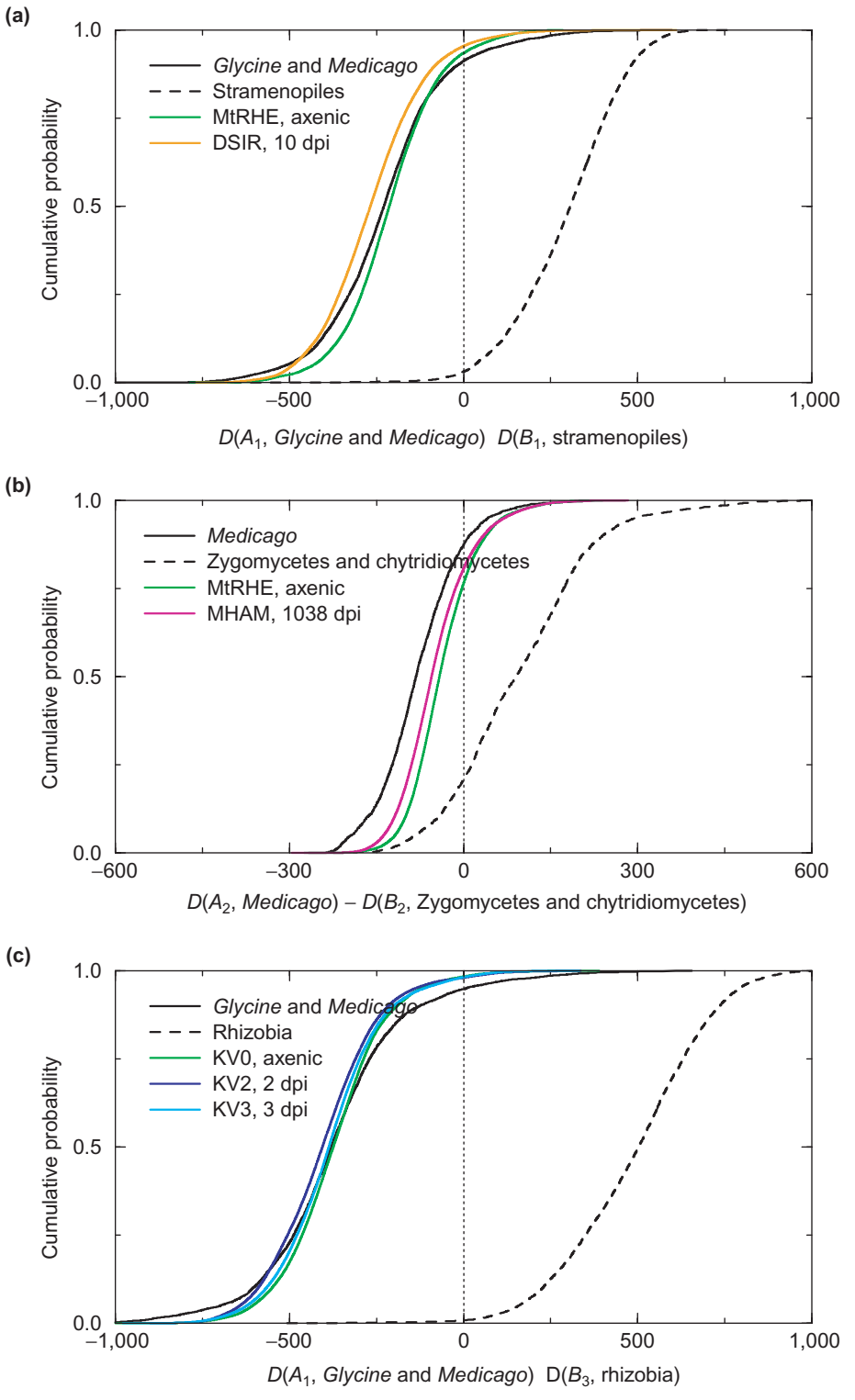

Figure 4
Dissimilarity distributions from *Medicago truncatula* libraries. Calibration curves compare plant training sets (A_1 and A_2, solid black lines) with one of three microbial symbiont training sets (broken black lines): **(a)** Stramenopile and *P. infestans* EST sequences (B_1); **(b)** pooled zygomycete and chytridiomycete coding sequences (B_2); and **(c)** sequences from the genera *Rhizobium*, *Sinorhizobium* and *Bradyrhizobium* (B_3). Cumulative distributions of test results from *M. truncatula* axenic and microbial symbiont mixed cultures appear in each panel (colored lines).

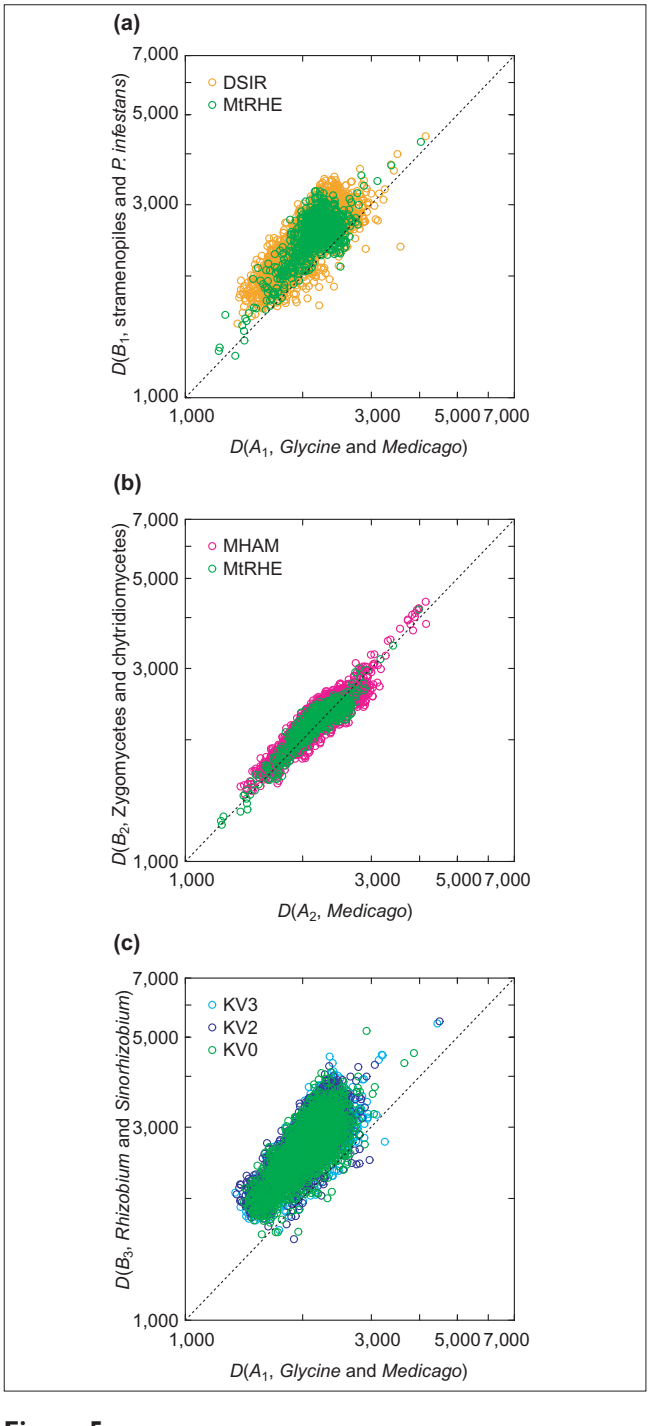

Figure 5
Paired comparison results from pure and mixed-culture *M. truncatula* libraries. Each point indicates the dissimilarity of a test sequence compared with a plant training set (A_1 or A_2) and one of three microbial symbiont training sets: **(a)** Stramenopile and *P. infestans* EST sequences (B_1); **(b)** pooled zygomycete and chytridiomycete coding sequences (B_2); and **(c)** sequences from the genera *Rhizobium*, *Sinorhizobium* and *Bradyrhizobium* (B_3). Sequences from *M. truncatula* axenic (green) and microbial symbiont mixed culture libraries are represented in each panel. The identity function ($y = x$) is also shown.

For mycorrhizal root libraries, we might explain the relative lack of symbiont sequences as resulting simply from a relative lack of transcripts in the host tissue. Most of the biomass in mycorrhizal roots is plant biomass [33]. We might therefore expect that most of the transcripts therein originate from the plant host. Confounding this result, the error rates in this comparison are the greatest among all the comparisons we performed, most likely because the evolutionary distance between fungi (zygomycetes and chytridiomycetes) and plants is the least among comparisons [34]. Also, zygomycete protein-coding sequences are rare in GenBank, which resulted in a small training set for these fungi, and may have amplified any biases. The high false-negative rate probably led to a failure to detect some symbiont transcripts.

In nodulating root libraries, we do not expect to observe an abundance of bacterial transcripts, because bacteria generally do not form polyadenylated mRNAs [35]. As the protocols used to extract and purify mRNAs from tissue lysate for the libraries cited in this study all relied on the presence of polyadenylation sites, we generally do not expect to find bacterial transcripts.

The abundance of putative microbial symbiont transcripts among sequences from a pure plant root library is difficult to interpret. The predicted portion of microbial transcripts was greater in the axenic root-hair enriched library than in mixed cultures. Error rates were greatest for comparisons between training sets from plant and pooled zygomycete and chytridiomycete sequences. Other than providing an 87% confidence level, the 13% false-positive rate does not completely explain why about 15% of root-hair enriched transcripts resemble fungal hexamer composition more closely than plants, and warrants further study.

Care had been taken to avoid contaminating plant tissue cultures by culturing seedlings in covered plates. Because of concern that ethylene accumulation in covered plates could improperly stimulate nodulation-related gene expression, seedlings were treated with Ag_2SO_4, an inhibitor of the plants' response to ethylene [6]. Inhibition of the ethylene response could have resulted in synthesis of transcripts that are uncharacteristic of plant roots. Analysis of another axenic root-hair enriched library, particularly one provided a carbon source to identify potential contaminants, and not treated with an inhibitor of ethylene response, would be an informative test.

These observations warrant further experimental scrutiny. The transcripts identified as most and least like plant or symbiont might also be studied in more detail as candidate participants in symbiosis. Symbiotic interactions, whether pathogenic or mutualistic, present novel challenges to both plant hosts and the biologists who study them. Computational approaches, in concert with experimental verification, can help resolve these challenges.

Methods and materials

Training sequences

Calibration

To characterize hexamer frequencies in plant hosts and their microbial symbionts, we collected sets of training sequences from public databases and edited them for quality. Training sets were chosen to be representative of, but obtained independently from, taxa participating in symbiotic associations for which a diagnosis of origin would be made. Because the species being compared are represented unevenly in public sequence databases, taxa were chosen so that roughly the same number of genes were analyzed in each training set, rather than simply to maximize the numbers of species or sequences present.

Training sets represent protein-coding sequences from three taxonomic groupings: plants (A_1, *Medicago* and *Glycine* spp.), either fungi (B_2, zygomycetes and chytridiomycetes) or stramenopiles (B_1), including ESTs from *P. infestans* [16], and bacteria (B_3, *Rhizobium*, *Sinorhizobium* and *Bradyrhizobium*). We performed pairwise comparisons with two different, taxon-specific training sets (A and B) to infer the origin of a transcript.

Training sets were obtained by querying the GenBank database using the Entrez retrieval tool [29-31]. A preliminary query by taxon name obtained all available nucleotide sequences from that taxon, then the Limits option excluded ESTs, STSs (sequence-tagged sites), GSSs (genome survey sequences), working draft sequences, and patented sequences from the query set. Organellar (mitochondrial and chloroplast) DNA was also excluded via the Limits option. A query term to require that a sequence contain a protein-coding region (CDS) was also added, which excluded ribosomal and transfer RNA sequences. The results consisted of all sequences that contain a nuclear protein-coding sequence available for that taxon at the time of the query. This was done on two separate occasions: in April and October 2000. (Changing slightly the composition of training sets between those dates did not notably affect the experimental outcome.)

Following a previously established protocol [19], we used a resampling procedure to evaluate the degree of overlap between distributions of hexamer composition obtained from comparing two training sets. In this protocol, we resampled each training set 40 times by random partitioning into training (for hexamer counts) and test calculation pools. To control for any bias introduced by length variation, a program randomly clipped 300 nucleotide fragments for word counting. As a result, one random 300 nucleotide fragment from each training sequence was present in the training set during a single resampling replicate; independent replicates contained different, randomly chosen training sequences and 300 nucleotide fragments. Values of the test statistic from 40 resampled replicates were pooled for calibration purposes.

As with the original protocol [19], we pooled the resulting test statistic distributions, normalized them as cumulative distributions, and then evaluated them for overlap. We call the resulting comparisons 'calibration curves', as they are not used directly to make inferences, but rather indirectly, to evaluate the degree of separation in hexamer counts from different taxa. Overlap of calibration curves should be minimal to yield the most statistically powerful results possible.

Due to considerable overlap of calibration curves between taxonomically general, inclusive training sets (that is, all eudicots, all fungi and miscellaneous eukaryotes, and all eubacteria, data not shown), we opted to work with specific training sets that included only the most species-specific sequences available, while maintaining approximately equal sample sizes across taxa.

The most challenging case was that of the arbuscular mycorrhizal fungi, for which very few protein-coding sequences are available. To increase the amount of data in this training set (B_2) without biasing sample sizes, we pooled sequences from all species in the zygomycetes with all available chytridiomycete coding sequences, and compared this training set with a set from a single plant genus, *Medicago* (A_2). We chose this option, rather than including an arbitrary subset of sequences from the ascomycetes and basidiomycetes, because zygomycetes and chytridiomycetes have diverged from their common ancestor less recently than the ascomycetes and basidiomycetes, based on 18S ribosomal RNA sequence data [34]. That is, the ascomycetes and basidiomycetes are more highly derived from the common fungal ancestor than zygomycetes and chytridiomycetes, which resemble more closely the ancestral state in modern lineages.

Data quality

Starting with a full set of sequences, we filtered for high-quality sequences by trimming regions having extensive ambiguous bases (N-rich) and poly(A) or poly(T) regions. The test statistic can be sensitive to the abundance of a single word [19]. Thus, we trimmed poly(A) and poly(T) sites to minimize the cases in which a test sequence resembles one training set more closely than the other, simply by virtue of having an abundance of the hexamer AAAAAA or TTTTTT. Similarly, test results obtained from short or N-rich sequences can be difficult to interpret [19]. We allowed no more than one N per hexamer and trimmed poly(A) or poly(T) tracts longer than 13 nucleotides. To accommodate for possible sequence chimeras, those sequences found to contain an internal poly(A) or poly(T) segment longer than 13 nucleotides were partitioned into two fragments, and the longer of the two fragments was used in analysis, provided its length was at least 300 nucleotides.

After trimming, we screened all remaining sequences of 300 nt or longer for similarity to *Escherichia coli* using BLASTN [20,21]. All BLAST searches used default parameters and

low-complexity filtering with the programs DUST or SEG. The decision to exclude non-coding RNA sequences from training sets was informed by the appearance of bimodal distributions of hexamer frequencies and a large degree of overlap between calibration curves (data not shown), likely a result of divergent evolutionary rates between protein-coding and non-coding sequences [36,37]. Chloroplast and mitochondrial sequences were eliminated to avoid complications due to variation in codon usage between nuclear and organellar genomes.

Table 2 summarizes counts of sequences and nucleotides in training sets before and after trimming and screening. All training sets obtained using the procedure described above are available as additional files.

Validation

To test the validity of word counting as a solution to the problem, we identified a set of 50 gene sequences from plants (*M. truncatula* and *G. max*), oomycetes (*Phytophthora*), zygomycetes (*Glomus versiforme*), and bacteria (*Sinorhizobium meliloti* and *Agrobacterium tumefaciens*), for which the function and origin have been characterized experimentally. We chose genes known to play a role in plant-microbe interactions, as well as genes that are found across taxa. We withheld these sequences, and partial transcripts of the same genes, from training sets prior to comparative lexical analysis, and calculated hexamer dissimilarities for each of the three training sets as described below.

Test sequences

To diagnose the species of origin for sequences expressed in symbiotic cultures, we collected sequences generated by distinct EST sequencing projects from the GenBank database [29-31]. Sequences from pathogenic interactions originated from cultures of a species from the genus *Phytophthora* with its plant host, such as *P. sojae* and soybean (*G. max*) isolated from inoculated hypocotyls two days after infection [18] and *P. medicaginis* and *M. truncatula* isolated from infected roots 10 days after infection (C. Vance, unpublished data). Sequences expressed during mutualistic interactions were obtained from cultures with *M. truncatula* and mycorrhizal (*Glomus versiforme*; M.J. Harrison, unpublished data) or rhizobacterial (*S. meliloti*; K. VandenBosch, unpublished data) endosymbionts several days after inoculation. Sequences expressed in pure, axenic cultures from *P. sojae* mycelia and zoospores [18] and from sterile, uninoculated *M. truncatula* roots [6] provided a basis for comparison in which no foreign transcripts were expected.

To maximize the reliability of diagnostic comparisons, we screened test sequences for high quality as for training sequences, and for low similarity to *E. coli*, chloroplast and mitochondrial genes, and non-coding RNA transcripts (ribosomal and transfer RNAs). Independent BLASTN comparisons identified sequences having very high similarity ($E < 10^{-100}$) to vector sequences or moderately high similarity ($E < 10^{-20}$) to non-nuclear or non-coding sequences obtained from GenBank. Sequences so identified were withheld from analysis. A summary of test sequences appears in Table 3. All test sequences obtained using the procedure described above are available as additional files.

Base content

We wrote a PERL program (countGC.pl) that calculates the GC base content of a sequence as the portion of guanine and cytosine residues among all unambiguous (non-N) nucleotides in a sequence. The hist method in R, version 1.1.1

Table 2

Training sets

Taxon	Raw		Trimmed		Screened	
	n	nt	n	nt	n	nt
Glycine	892	1,265,829	834	1,219,114	826	1,184,951
Medicago (A_2)	401	561,104	382	519,739	380	513,868
Total, plants (A_1)					1,206	1,698,819
Stramenopiles	199	299,113	184	287,600	181	279,900
P. infestans	2,131	1,219,463	2,102	1,209,113	2,082	1,199,372
Total, stramenopiles (B_1)					2,263	1,479,272
Zygomycetes	232	343,817	212	329,222	211	327,229
Chytridiomycetes	82	123,698	78	119,754	78	119,754
Total, Fungi (B_2)					289	446,983
Rhizobium	478	1,430,132	444	1,404,883	444	1,404,883
Sinorhizobium	320	900,294	312	898,687	312	898,687
Bradyrhizobium	153	471,309	146	465,307	146	465,307
Total, rhizobacteria (B_3)					902	2,768,877

Number of sequences (n) and nucleotides (nt), as raw, trimmed (removed N-rich regions, poly(A) and poly(T) sites), and screened sequences (removed ribosomal, chloroplast, and mitochondrial DNA and remaining sequences shorter than 300 nucleotides).

[38] aggregated continuous percentages into discrete histogram bins, using bin sizes of 2% difference in GC, with inclusive lower bin boundaries and exclusive upper bounds; the lm method tested for linear correlation of the dissimilarity test statistic t with GC.

Comparative lexical analysis

White *et al.* [19] used a likelihood-ratio test to determine whether word frequencies from a particular sequence more closely resemble the frequency distribution of control data sets from the taxon being sequenced or a distantly related outgroup. They computed a test statistic $t(A,B,x)$ for each sequence x as the difference of log-likelihood ratio dissimilarity measures, $D(A,x) = -2\log\lambda(A,x)$, for two data sets, a control set A and an outgroup B, such that $t(A,B,x) = D(A,x) - D(B,x)$. A negative value for t indicates that the sequence more closely resembles words from A; conversely, a positive value indicates a likely contaminant related to B. (Dissimilarity is conceptually related to distance. However, dissimilarity does not measure distance because it does not possess the mathematical properties of a distance metric [39].) Unlike the calculation of calibration curves, in which 300-nucleotide subsequences are randomly resampled, hexamer dissimilarity is measured over the whole length of a test sequence when inferring a transcript's origin. Originally, the investigators used the null hypothesis that no difference exists for dissimilarity measures between the two data sets, or that $t(A,B,x) = 0$ [19]. White *et al.* [19] tested two alternative hypotheses: that $t < 0$, being more like A, or $t > 0$, like B.

Lexical analysis using pentamers or heptamers yields similar error rates and very highly correlated values for the test result (not shown). Because White *et al.* [19] reported the best results were obtained using hexamers, and because a word size of six nucleotides corresponds to the size of a dicodon, we chose to analyze hexamer frequencies. To use longer words requires more training data, because the number of possible words increases exponentially with increasing word size. Use of shorter words may be adequate for some applications and will be investigated in future work.

Though we used White's word-counting methods, we did make slight modifications. We simplified one program (called hybridize) to compute individual dissimilarity values, rather than paired differences; a patch that details how to modify the C program is available (see hyb2dis.txt in additional data files online). More importantly, we amended the null hypothesis and interpreted calibration curves to test for statistically significant dissimilarity differences. Though the likelihood-ratio test statistic indicates the magnitude of similarity to A or B, we do not know what values for t are significant with known confidence. When testing hypotheses, one can make two types of error: type I, or false positives, and type II, false negatives [28]. The false-positive rate is denoted α and false-negative rate β. We determine α and β from overlap in the calibration curves. Inferring error rates from calibration curves is justified because we know the correct answer and determine the error rate via resampling, as with bootstrap methods to infer error rates or confidence intervals [40].

We are interested in knowing from which of two organisms a sequence originated, and are reasonably confident that it came from either one or the other. Thus, we assume it came from one and test whether we have evidence to refute this assumption. The null hypothesis here is that sequence x is from A. Alternatively, it might be from B. Evaluating the calibration curve overlap at $t = 0$ quantifies the associated error rates. The cumulative distribution function (*cdf*) of taxon B specifies β where cdf_B intersects 0; the *cdf* from A specifies α as $1-cdf_A(0)$. We can thus resolve the problem with known confidence P: $P(t > 0) = \alpha$. All other computations were performed as described previously [19]. Software used for

Table 3

Test sets

Species	Tissue	Library (ID)	Raw		Trimmed		Screened	
			n	nt	n	nt	n	nt
P. sojae	Mycelia	MY	969	527,295	902	510,010	895	506,086
P. sojae	Zoospores	ZO	1,013	583,520	960	569,576	957	567,976
+ *G. max*	2 dpi	HA	994	577,626	938	563,226	927	556,305
M. truncatula	Root hairs	MtRHE	899	539,719	893	536,787	890	534,037
+ *G. versiforme*	10-38 dpi	MHAM	3,259	1,785,721	3,030	1,735,390	3,017	1,725,491
+ *P. medicaginis*	10 dpi	DSIR	2,462	1,324,815	2,289	1,287,568	2,284	1,282,518
M. truncatula	Roots	KV0	2,718	1,387,832	2,550	1,351,137	2,492	1,318,131
+ *S. meliloti*	1 dpi	KV1	1,125	562,452	1,012	537,644	1,003	531,813
+ *S. meliloti*	2 dpi	KV2	1,960	976,344	1,732	926,953	1,726	922,433
+ *S. meliloti*	3 dpi	KV3	2,375	1,316,430	2,217	1,279,691	2,173	1,251,795

Number of EST sequences (n) and nucleotides (nt) as raw, trimmed (limited lengths of N-rich regions, poly(A) and poly(T) sites), and screened (removed ribosomal, chloroplast, and mitochondrial DNA, and remaining sequences shorter than 300 nt) sequences. Transcripts were isolated from the cDNA library indicated by the ID column. dpi, days post-inoculation, indicating mixed plant-microbe cultures.

lexical analysis was obtained via anonymous ftp from the TIGR software FTP site [41].

Acknowledgements

We thank Callum Bell, Mark Gijzen, Maria Harrison, Tom Kepler, Deb Samac, and Bruno Sobral for valued discussions and feedback. Comments from B. M. Tyler and an anonymous reviewer on an earlier version of this work greatly enhanced its presentation. PTH thanks the Santa Fe Institute for support and inspiration.

Additional data files

The following files are available for download with the online version of this article:

countGC.pl: PERL script used to compute GC content of sequences analyzed.

hyb2dis.txt: patch file that converts White's hybridize program to compute individual dissimilarity values.

Training sets (GlycineMedicago.txt, Rhizobia.txt, Stramenopiles.txt, ZygoChytrid.txt): FASTA-formatted text files that contain the sequences used for calibration and comparison.

Test sets (PsojaeHA.txt, PsojaeMY.txt, PsojaeZO.txt, MtRHE.txt, DSIR.txt, MHAM.txt, KV0.txt, KV2.txt, KV3.txt): FASTA-formatted text files containing transcripts analyzed, edited for quality.

Test results (PsojaeHA.dat, PsojaeMY.dat, PsojaeZO.dat, MtRHE-A1B1.dat, MtRHE-A2B2.dat, DSIR.dat, MHAM.dat, KV0.dat, KV2.dat, KV3.dat): text files that contain transcript analysis results, sorted from least to most plant-like.

References

1. Adams MD, Fields C, Venter JC: *Automated DNA sequencing and analysis*. London: Academic Press, 1994.
2. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656.
3. Cook DR: ***Medicago truncatula* - a model in the making!** *Curr Opin Plant Biol* 1999, **2:**301-304.
4. Rounsley S, Lin X, Ketchum KA: **Large-scale sequencing of plant genomes.** *Curr Opin Plant Biol* 1998, **1:**136-141.
5. Somerville C, Somerville S: **Plant functional genomics.** *Science* 1999, **285:**380-383.
6. Covitz PA, Smith LS, Long SR: **Expressed sequence tags from a root-hair-enriched *Medicago truncatula* cDNA library.** *Plant Physiol* 1998, **117:**1325-1332.
7. Hillier L, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, *et al.*: **Generation and analysis of 280,000 human expressed sequence tags.** *Genome Res* 1996, **6:**807-828.
8. Höfte H, Desprez T, Amselem J, Chiapello H, Rouze P, Caboche M, Moisan A, Jourjon M, Charpenteau J, Berthomieu P, *et al.*: **An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA from *Arabidopsis thaliana*.** *Plant J* 1993, **4:**1051-1061.
9. Nelson MA, Kang S, Braun E, Crawford ME, Dolan PL, Leonard PM, Mitchell J, Armijo AM, Bean LL, Blueyes E, *et al.*: **Expressed sequences from conidial, mycelial, and sexual stages of *Neurospora crassa*.** *Fungal Genet Biol* 1997, **21:**348-363.
10. Newman T, de Bruijn F, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, *et al.*: **Genes galore: a summary of the methods for accessing the results of large scale partial sequencing of anonymous *Arabidopsis thaliana* cDNA clones.** *Plant Physiol* 1994, **106:**1241-1255.
11. Bortoluzzi S, Danieli G: **Towards an *in silico* analysis of transcription patterns.** *Trends Genet* 1999, **15:**118-119.
12. Harrison MJ: **Biotrophic interfaces and nutrient transport in plant/fungal symbioses.** *J Exp Bot* 1999, **50:**1013-1022.
13. Birch P, Avrova A, Duncan J, Lyon G, Toth R: **Isolation of potato genes that are induced during an early stage of the hypersensitive response to *Phytophthora infestans*.** *Mol Plant-Microbe Interact* 1999, **12:**356-361.
14. Györgyey J, Vaubert D, Jiménez-Zurdo JI, Charon C, Troussard L, Kondorosi A, Kondorosi E: **Analysis of *Medicago truncatula* nodule expressed sequence tags.** *Mol Plant-Microbe Interact* 2000, **13:**62-71.
15. Harrison MJ: **Molecular and cellular aspects of the arbuscular mycorrhizal symbiosis.** *Annu Rev Plant Phys Plant Mol Biol* 1999, **50:**361-389.
16. Kamoun S, Hraber PT, Sobral BWS, Nuss D, Govers F: **Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences.** *Fungal Genet Biol* 1999, **28**:94-106.
17. van Buuren ML, Maldonado-Mendoza IE, Trieu AT, Blaylock LA, Harrison MJ: **Novel genes induced during an arbuscular mycorrhizal (AM) symbiosis formed between *Medicago truncatula* and *Glomus versiforme*.** *Mol Plant-Microbe Interact* 1999, **12:**171-181.
18. Qutob D, Hraber P, Sobral B, Gijzen M: **Comparative analysis of expressed sequences in *Phytophthora sojae*.** *Plant Physiol* 2000, **123:**243-254.
19. White O, Dunning T, Sutton G, Adams M, Venter JC, Fields C: **A quality control algorithm for DNA sequencing projects.** *Nucleic Acids Res* 1993, **21:**3829-3838.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
22. Smith T, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
23. Braun E, Halpern A, Nelson M, Natvig D: **Large scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*.** *Genome Res* 2000, **10:**416-430.
24. Knight RD, Freeland SJ, Landweber LF: **A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.** *Genome Biol* 2001, **2:**research0010.1-0010.13.
25. Grantham R, Gautier C, Gouy M: **Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type.** *Nucleic Acids Res* 1980, **8:**1893-1912.
26. Li WH, Graur D: *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates, 1991.
27. Dunning T: **Accurate methods for the statistics of surprise and coincidence.** *Comp Linguistics* 1993, **19:**61-74.
28. Freund JE, Walpole RE: *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
29. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2000, **28:**15-18.
30. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2000, **28:**10-14.
31. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29:**11-16.
32. Eddy S: **Noncoding RNA genes.** *Curr Opin Genet Dev* 1999, **9:**695-599.
33. Toth R, Miller RM, Jarstfer AG, Alexander A, Bennet EL: **The calculation of intraradical fungal biomass from percent coloniza-**

tion in vesicular-arbuscular mycorrhizae. *Mycologia* 1991, **83:**553-558.
34. van de Peer Y, de Wachter R: **Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA.** *J Mol Evol* 1997, **45:**619-630.
35. Lewin B: *Genes V*. Oxford, UK: Oxford University Press, 1995.
36. Futuyma DJ: *Evolutionary Biology* Third edition. Sunderland, MA: Sinauer Associates, 1998.
37. Harvey P, Pagel M: *The Comparative Method in Evolutionary Biology*. Oxford UK: Oxford University Press, 1991.
38. Ihaka R, Gentleman R: **R: a language for data analysis and graphics.** *J Comp Graphic Stat* 1996, **5:**299-314.
39. Weir BS: *Genetic Data Analysis* Second edition. Sunderland, MA: Sinauer Associates, 1996.
40. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall, 1993.
41. **TIGR Software** [ftp://ftp.tigr.org/pub/software/qc]