Funct Integr Genomics (2003) 3:33–38
DOI 10.1007/s10142-002-0075-1

REVIEW

Bryan Clarke · Mark Lambrecht · Seung Y. Rhee

Arabidopsis genomic information for interpreting wheat EST sequences

Received: 12 April 2002 / Accepted: 7 July 2002 / Published online: 5 September 2002

Abstract The resources available from *Arabidopsis thaliana* for interpreting functional attributes of wheat EST are reviewed. A focus for the review is a comparison between wheat EST sequences, generated from developing endosperm tissue, and the complete genomic sequence from *Arabidopsis*. The available information indicates that not only can tentative annotations be assigned to many wheat genes but also putative or unknown *Arabidopsis* gene annotations can be improved by comparative genomics.

Keywords Wheat · *Arabidopsis* · Comparative genomics

Introduction

The complete genomic sequence from *Arabidopsis* provides many benefits for functional and evolutionary biology. Genomic projects often focus on model species to provide detailed information for use by researchers working with other species. For plant biologists the models are *Arabidopsis* and rice. Although rice is the model for cereals, the genome sequence of *Arabidopsis* is closer to completion and is more clearly annotated, and thus provides the best opportunities for assigning possible functions to new genes. In addition, *Arabidopsis* is continuing to provide a focus for genome and proteome analyses, and aspects of plant development in publicly funded research programs.

B. Clarke (✉)
CSIRO Plant Industry PO Box 1600, ACT 2601, Australia
e-mail: Bryan.Clarke@csiro.au

M. Lambrecht · S.Y. Rhee
The Arabidopsis Information Resource,
Carnegie Institution of Washington, Department of Plant Biology,
260 Panama Street, CA 94305, USA

M. Lambrecht
Centre of Microbial and Plant Genetics, University of Leuven,
Kasteelpark Arenberg 20, 3001 Leuven, Belgium

Sequencing has shown that the genomes of eukaryotes contain a large fraction of conserved sequences, especially those genes coding for the core biological functions of the cell. In general, cell functions such as assembly of the cytoskeleton, or essential processes such as DNA replication, repair, recombination and metabolism, as well as cell division, protein synthesis and vesicle trafficking are largely preserved among all eukaryotes. However, genes involved in regulatory processes, such as signal transduction and transcriptional regulation, can be quite diverse (Poethig 2001). Comparative analysis of predicted proteins has shown that nearly 20% of *Drosophila melanogaster* proteins have putative orthologues in both *Caenorhabditis elegans* and *Saccharomyces cerevisiae* (Rubin et al. 2000).

The *Arabidopsis* genome is estimated to contain approximately 25,000 genes encoding approximately 11,000 unique proteins (The *Arabidopsis* Genome Initiative 2000). Comparisons between the dicots *Brassica napus* and *Arabidopsis* showed that, on average, there is 87% sequence conservation in the protein coding sequences and extensive colinear gene order (Cavell et al. 1998). Despite the monocot-dicot divergence 200 (±40) million years ago (Wolfe et al. 1989), comparison between rice and *Arabidopsis* genes suggest substantial sequence identity. For example, 58% of non-singleton ESTs grouped into EST clusters shared between *Arabidopsis* and rice (Ewing 1999) and 64 randomly chosen *Arabidopsis* and rice putative orthologous genes showed 30–100% identity with a peak at 75% (Somerville and Somerville 1999). Some conservation of genome microstructure is also evident between monocot and dicot species. At the level of intron-exon structure, conservation is observed between cereals and *Arabidopsis* (Dubcovsky et al. 2001; reviewed in Appels et al. 2002, this issue). Mayer et al. (2001) identified 56 genes on a 340-kb genome segment of rice, and of these 22 genes had sequence conservation in *Arabidopsis,* with at least 5 genes showing conserved gene order on each of four genome segments. However, observations by Gale and Devos (2001) indicate that, although some colinear gene order

is retained between *Arabidopsis* and rice, the utility of predictive positional cloning between the two models is likely to be minimal.

To determine the function of an unknown gene, it is often necessary to create a null mutation of that gene and characterize the change in phenotype. However, it may be possible to gain insight about the gene of interest by a comparative genomic approach, where orthologous genes in model plants are identified for which more information may be available. The extent of gene duplication found in *Arabidopsis* and in other plants and the rate of genome sequencing efforts in general, indicate that reverse genetics is likely to become the more common first approach in deciphering a gene function. Therefore, the ability to determine which genes in the crop of interest may have functional homology to a model plant is important. True orthologous genes can be assigned only if the genes share the same chromosomal location in addition to complete sequence identity (Bork et al. 1998). This strict definition is clearly not possible using *Arabidopsis* as the model plant for wheat. Nevertheless, a high degree of overall pair-wise similarity between genes in a completely sequenced organism and a distantly related species could be used as a first step to identify putative orthologues and be the basis for further analysis to determine whether they share the same or similar function. To date only a fraction of the sequenced genes have been studied experimentally in any organism, and often a functional annotation is assigned to a gene on the basis of its sequence similarity to another. In addition, the relationship between sequence similarity and protein function for most proteins has not been established. These problems are particularly acute with database searches using EST sequences because, often, only limited regions of the genes are being compared. Given these caveats, any attempt at making functional annotations based on sequence similarity should be carried out with clear descriptions of criteria and parameters used for such analyses. A set of criteria that should be considered in functional annotations of wheat genes based on sequence similarity to *Arabidopsis* are developed in this review.

A starting point

A set of wheat endosperm EST sequences that was used as a starting point for relating wheat sequences to those of the *Arabidopsis* genome was described in Clarke et al. (2000) (GenBank accession number BQ 605537–609969). The error rate for the sequence data is <2% N (unresolved nucleotides) over 500 bp. The *Arabidopsis* protein and nucleotide sequence databases were from TIGR (http://www.tigr.org/tdb/e2k1/ath1/).

Clustering of EST sequences is most commonly carried out using the PHRAP program (Green 2002) using the default settings of the program. This was the procedure carried out for the 4,433 wheat ESTs used as a focus in this paper. Two clustering steps, the first to cluster EST sequences into contig sequences using a self BLAST approach, and the second to further combine contig sequences using the PHRAP program, were used to create a non-redundant set of sequences that were systematically compared to *Arabidopsis* (ftp://arabidopsis.org/Genes/est_mapping/Wheat_comparison/BRY_clustered).

The extra PHRAP clustering step was introduced to remove redundancy as far as possible. This produced contig sequences that were on average longer and had a lower percentage of low complexity sequences than their constituent ESTs and therefore, on average, produced higher scoring alignments. The clustering process resulted in 1,348 unclustered sequence (singletons) ESTs and 789 clustered sequences (contigs). The resulting FASTA file, containing 2,137 unique sequences, were used for further analyses.

Comparing wheat and *Arabidopsis* gene sequences

Full EST comparison and BLAST result analysis continue to provide the most effective procedure for determining the level of sequence similarity and this was carried out for the analysis of the *Arabidopsis* genome sequence and the wheat endosperm clones. The wheat EST sequence collection was compared against the *Arabidopsis* protein data set generated by TIGR (http://www.tigr.org/tdb/e2k1/ath1/; January 2002 release, downloaded from TAIR – ftp://arabidopsis.org/home/tair/Sequences/blast_datasets/). The *Arabidopsis* intergenic sequences were derived from TIGR's January 2002 release (ftp://arabidopsis.org/home/tair/Sequences/blast_datasets/) and the total non-redundant, GenBank protein sequence database was from Benson et al. (2000; release 126.0, October 15 2001). Extensive use was made of the BLASTX programs (Altschul et al. 1997) with the standard input parameters (default BLOSUM62 matrix) coupled to a BLAST report parser (Bioperl BLAST module, see http://www.bioperl.org). Multiple alignments were generated with ClustalW 1.8 and graphically represented by the GeneDoc program version 2.6.002 and by the perl/GD programs module (http://stein.cshl.org/WWW/software/GD/), an interface to the GD graphics library. All programs were executed on a dual-processor workstation running under the Redhat Linux 7.0 operating system.

The 2,137 unique wheat EST sequences (containing 789 contigs and 1,348 singletons) were matched against the *Arabidopsis* protein sequence dataset from TIGR using BLASTX at the e-6 cut-off value. At this e-value cut-off, all wheat EST sequences that aligned for their entire length to the *Arabidopsis* protein sequence, as well as some with partial alignments (e.g. a functional domain), were included. The non-matching EST sequences (e-value above e-6) were not considered further in this comparison.

In the BLASTX comparison of wheat EST sequences to *Arabidopsis*, approximately 50% of the *Arabidopsis*

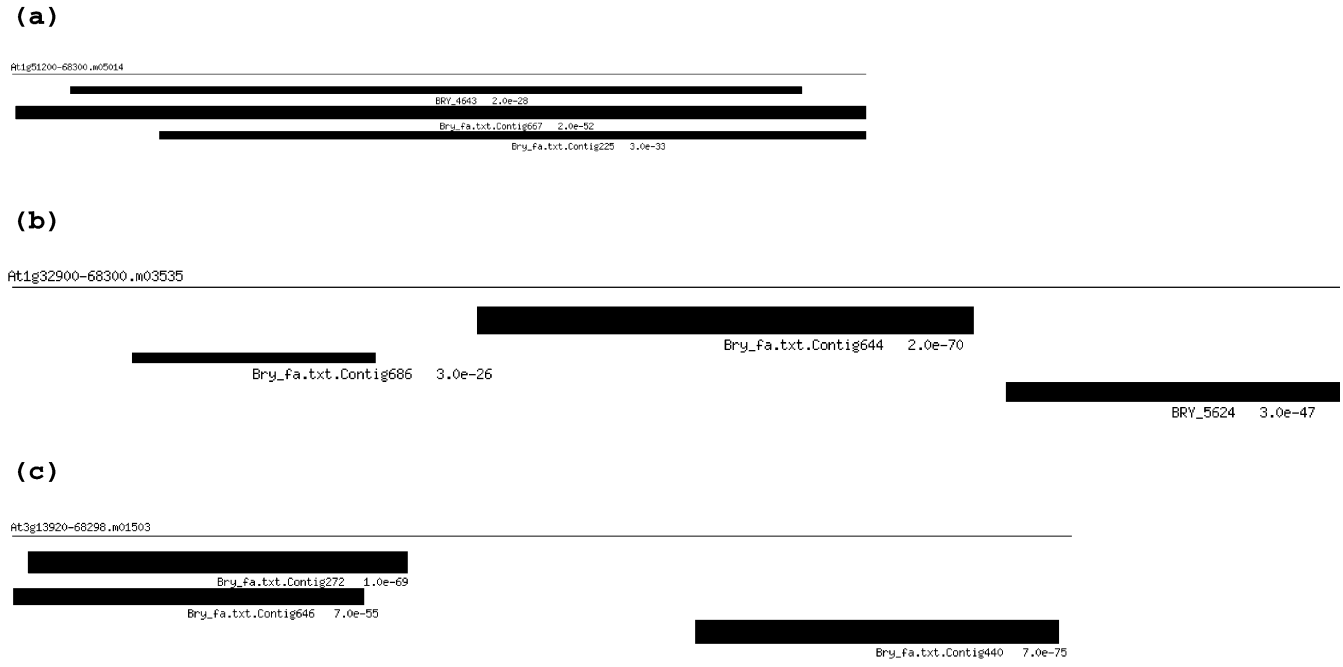

Fig. 1 Three different types of alignments of unique EST sequences that match to the same *Arabidopsis* protein sequence as a best hit: the alignments overlap each other (**a**), are non-overlapping (**b**), or contain both overlapping and non-overlapping ESTs (**c**). The *thickness of the EST bar* is proportional to the e-value in the BLAST output

genes matched two or more wheat ESTs. In these cases, where two or more wheat sequences matched one *Arabidopsis* gene, three different situations arise (see Fig.1). In about 64% of the cases (Fig. 1a) wheat EST sequences overlap, matching the same or an overlapping part of the *Arabidopsis* gene sequence. These could reflect gene duplications in wheat or remaining redundancy in the wheat EST sequences, despite extended clustering procedures. In 30% of the sequence comparison, the wheat EST sequences do not overlap and match different parts of the *Arabidopsis* gene sequence (Fig. 1b). These could point to insufficient sequence information to produce a contiguous sequence in the PHRAP alignment of the EST data set, or the matching *Arabidopsis* gene could be composed of a combination of functional domains. In a few cases, both of these two types of matching are shown, which probably result from a combination of the above two categories (Fig. 1c).

The 2,137 non-redundant wheat ESTs were finally grouped with 1,130 unique *Arabidopsis* genes (several of these had more than wheat sequence matching) and 681 sequences did not match at or matched above the e-6 cut-off value. Of the matched sequences, 853 were annotated with a gene description and 122 were listed as unknown and 155 as putative proteins. This information provides a first approximation for the gene function of around 75% of the matched wheat EST sequences. In addition, as the putative and unknown categories are functionally annotated in *Arabidopsis* they will provide information for the functional assignment of the wheat genes that they match, especially as 95% of these matches had an e-value below 10. The 1,130 unique genes were distributed over all five chromosomes in the *Arabidopsis* genome with 297 genes on chromosome I, 163 on chromosome II, 226 on chromosome III, 162 on chromosome IV and 282 on chromosome V.

In looking at these wheat sequences that group with one *Arabidopsis* gene, the combinatorial nature of proteins must be taken into account. It is possible that these wheat genes are different and that they just share a functional domain with the *Arabidopsis* sequence they match. This information is still valid, however, as it can place the wheat sequence into a broad family of genes that share this functional domain. The number of unique wheat genes in this data set is, therefore, likely to be larger than that predicted by this comparative genomic analysis.

The results reflect the value of *Arabidopsis* as a model for plants. In the case of the *Arabidopsis* genes the whole sequence is available and this provides a complete amino acid template of the gene to which the limited translated EST sequences can be compared. This comparison of the translated sequences provides a more accurate way of comparing distantly related gene sequences and allows the short sequence fragments from different parts of the same gene, provided by ESTs, to be grouped.

Arabidopsis functional information data mining

Gene ontology (GO) annotations for the *Arabidopsis* loci matching the wheat ESTs were retrieved using TAIR's GO annotation search (http://arabidopsis.org/tools/ bulk/go/). Lines carrying insertions in the *Arabidopsis* loci were re-

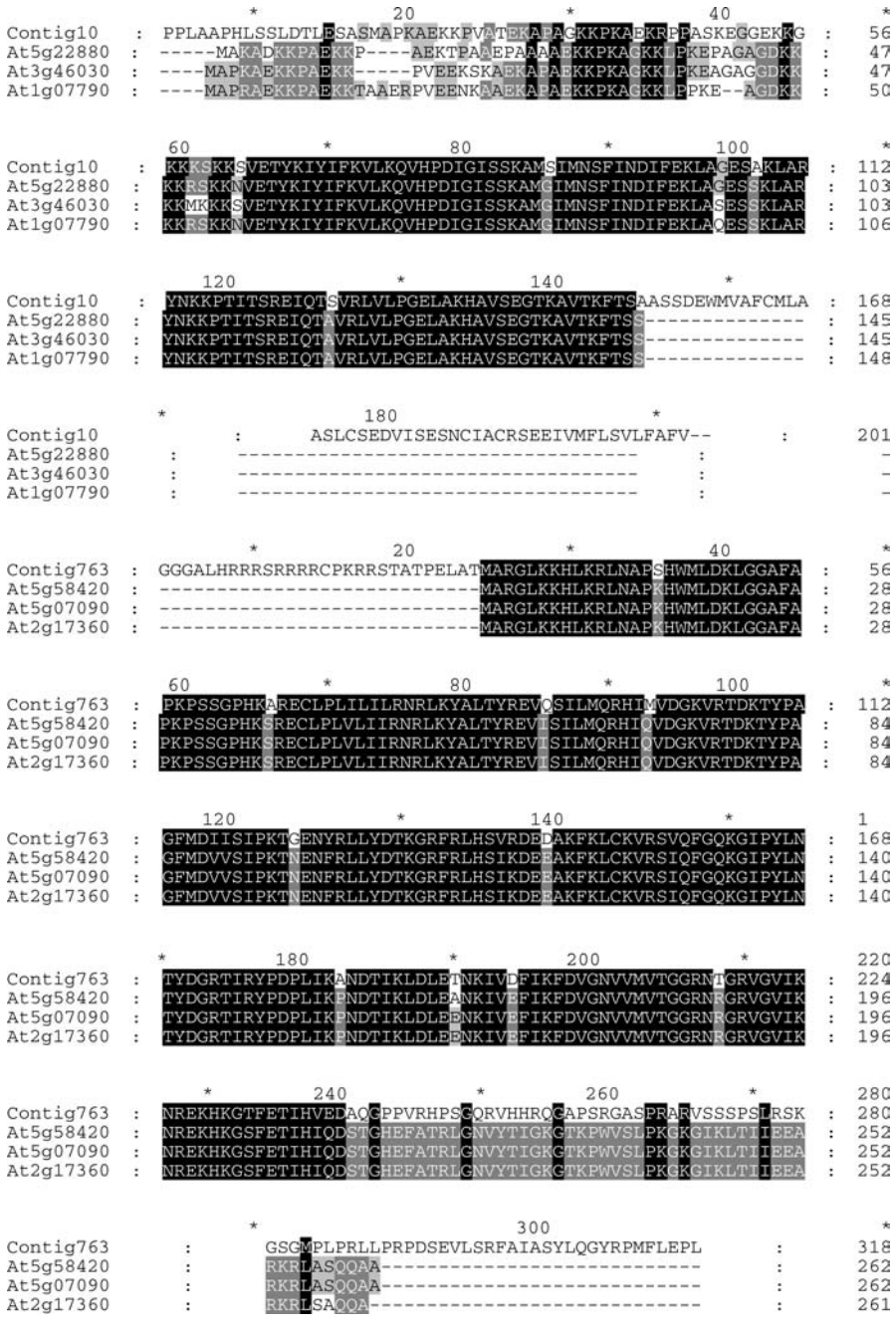

Fig. 2a, b ClustalW alignment of two wheat EST sequences with *Arabidopsis* sequences. *Shading* indicates similarity: *black* 100%, *dark gray* 75%, *light Grey* 50%

trieved using TAIR's Functional Genomics search (http://arabidopsis.org/cgi-bin/2010_projects/2010_search.pl). In order to interpret the data, the factors that would be important in assigning functional annotation from an automated sequence comparison of wheat EST and the *Arabidopsis* protein sequences using BLASTX were considered. A key result from a BLASTX analysis is the score of the alignments. The alignment score is affected by several factors, including the amino acid residue composition and length of the query sequences, the scoring matrix used and the total length of the database being searched. The region of similarity with regard to the query and subject sequence lengths should also be considered.

For example, two sequences, a histone and a ribosomal protein, aligned between wheat and *Arabidopsis* both show that there is strong sequence identity, (shown in Fig. 2). These alignments cover a significant portion of the *Arabidopsis* gene and it could be assumed they share functional homology. It is worth noting, in these comparisons, that the histone gene (Fig. 2a) is shorter than the S4 ribosomal protein gene (Fig. 2b) and the e-values for the matches were e-43 for the histone and e-115 for the ribosomal protein, respectively, indicating that the e-value is dependent on the length of the genes being compared. This indicates, that now we have fully annotated full-length sequences for genes in the model genomes,

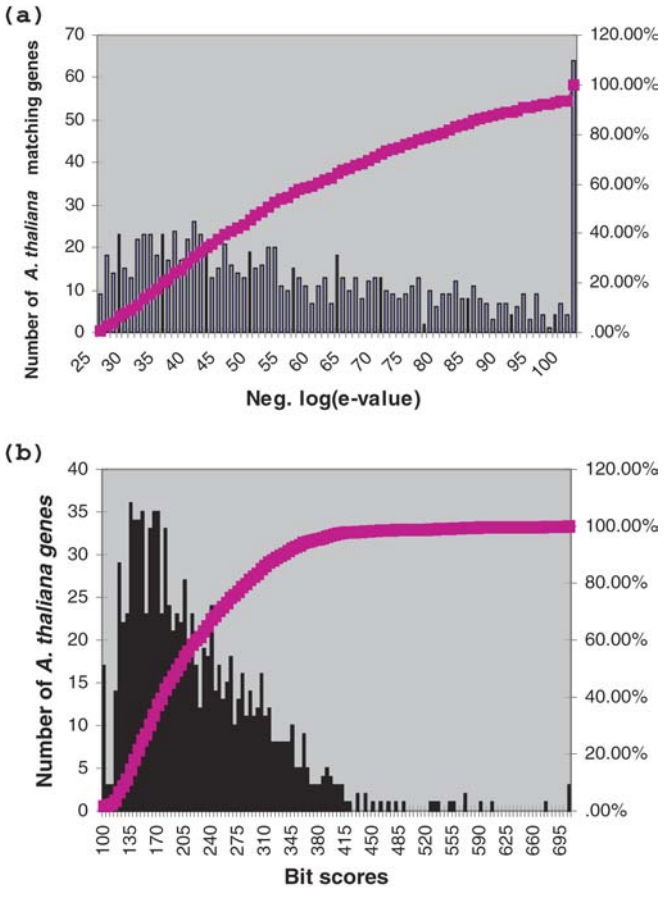

Fig. 3 Histogram of e-values (**a**), bit scores (**b**) in the BLASTX output of wheat EST unique sequences and *Arabidopsis* protein sequence comparison. The *curve* represents the cumulative percentage plot and can be used to read the percentage of *Arabidopsis* protein sequences that have a negative log e-value (**a**) or bit score (**b**) below or equal to the input value. Note: the EST sequences matching at an e-value of e-100 and lower were grouped in one bin, approximately 5–10% of the total matching sequences. The histograms of the BLAST reported e-value and bit scores were generated using Microsoft Excel

we need to develop more sophisticated BLAST algorithms that reflect, by way of a more significant e-value, a complete match to a short gene.

A histogram, representing the distribution of e-values and bit scores for the BLASTX comparison of the wheat EST sequences and the *Arabidopsis* protein sequence dataset, is represented in Fig. 3. Although the e-values were distributed more uniformly than the bit scores, both distributions show a similar pattern in their shape, with peaks centered around e-30 and e-45 and 140 and 180 for the e-value and bit scores, respectively (note that all e-values under e-100 have been binned in one category in Fig. 3a).

Functional annotation of wheat EST sequences

The matching *Arabidopsis* protein sequences and their descriptions, with BLAST scores below e-50, are available at TAIR (ftp://arabidopsis.org/home/tair/Genes/est_mapping/). Of the 853 annotated *Arabidopsis* protein sequences that matched wheat EST sequences as best hits, 580 have been assigned with gene ontology terms (Gene Ontology Consortium, http://www.geneontology.org, TAIR, http://arabidopsis.org.org/info/ontologies/). Most of these annotations belong to general cellular machinery (data not shown). Of the 853 protein sequences, 411 of them had T-DNA insertions, and lines carrying these insertions are available (TAIR, http://arabidopsis.org/info/2010_projects/). Among the 3,735 *Arabidopsis* protein sequences matching to the wheat ESTs (including hits other than the best hit), 1,627 of them have T-DNA insertions. Lines carrying these insertions are distributed by the Arabidopsis Biological Resource Center (ABRC) and can be ordered from TAIR (http://arabidopsis.org/servlets/SeedSearcher).

Conclusions

EST sequences are often the first genomic data available for an organism, and their annotation provides information about the genes these sequences represent. Although there is insufficient experimental information to infer function from structure, sequence analysis can facilitate the process of experimentally determining the function of proteins in a more efficient manner.

The information available for the *Arabidopsis* genome is a valuable resource because it provides the complete sequence of a flowering plant, including the multigene families and functional annotations for many of these genes. The rapid progress in functional characterization of the proteome of *Arabidopsis* will help to better define the relationship between structure and function of proteins in plants. It has been shown here that both the structural and functional annotation of the wheat ESTs can be improved from their sequence comparison. From the analysis, we found several types of matches that need to be considered when making functional annotation from sequence analysis of unknown sequences against the genome of a distantly related organism. The first case is where the wheat EST sequence covers the whole or a substantial part of an *Arabidopsis* protein. Second, the majority of the wheat EST sequence matches to an analogous region of a larger *Arabidopsis* protein. Third, the wheat EST has strong similarity across a short region of disparate *Arabidopsis* proteins. The level of sequence similarity at which the first two types of match is seen is generally with e-values of e-100 or lower. Most of the matches in the third category have e-values that are above e-100. The first two types of matches should get a higher confidence level for the functional annotation than the third category. For example, in the third category, an alignment of the full-length *Arabidopsis* genes, which match a wheat sequence with e-values between e-60 and e-113, the *Arabidopsis* protein sequences share sequence similarity across the common region with wheat but are clearly different outside this domain. The

different types of matches described here can be used to assign different levels of confidence in functional annotation, or can be used as parameters for a cut-off, in addition to the e-value, in an automatic functional annotation based on sequence. However, there are caveats, as with any computational methods, in using these factors to assign function. For example, proteins with completely different folds can have the same function (convergent evolution), but also proteins with the same folds can have different functions (divergent evolution; Hegyi and Gerstein 1999). This is not to say that proteins with strong but not complete sequence similarity do not have functional equivalence. Kleywegt (1999) suggests that a small number of peptide residues do the functional work of the protein, whereas the overall fold of the protein is important as an enabling framework. This may be the case in the alignments with several *Arabidopsis* EST sequences and one wheat sequence; the completely conserved stretches may be the functional domains and therefore the proteins may still have functional equivalence.

Using *Arabidopsis,* as indeed any other "model organism", to determine gene function from a plant of interest based on computational methods alone is very risky and, as outlined above, these comparisons must be made with explicit descriptions of what parameters were used in making the annotations. As expected, the majority of the wheat sequences with good similarity to *Arabidopsis* largely fall into those categories associated with the core cellular machinery. The comparison to a distantly related species can thus be considered as a "filter", to distinguish and set aside wheat EST sequences that are more likely to be involved in species-specific functions and therefore pursued further with experimentation.

Acknowledgements M.L. is a postdoctoral fellow of the Fund for Scientific Research Flanders (F.W.O.-Vlaanderen) and of the Carnegie Institution of Washington. S.Y.R. is funded by NSF grant number DBI-9978564.

References

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

Appels R, Francki M, Chibbar R (2002) Advances in cereal functional genomics. Funct Integr Genomics (this issue)

Arabidopsis Genome Initiative. The (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000) GenBank. Nucleic Acids Res 28:15–18

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. J Mol Biol 283:707–725

Cavell AC, Lydiate DJ, Parkin IAP, Dean C, Trick M (1998) Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated region within the *Brassica napus* genome. Genome 41:62–69

Clarke BC, Hobbs M, Skylas D, Appels R (2000) Genes active in developing wheat endosperm. Funct Integr Genomics 1:44–55

Dubcovsky J, Ramakrishna W, San Miguel P, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol 125:1342–1353

Ewing R, Poirot O, Claverie JM (1999) Comparative analysis of the *Arabidopsis* and rice expressed sequence tag (EST) sets. In Silico Biol 1:197–213

Gale MD, Devos K (2001) Comparative genetics and cereal evolution. Isr J Plant Sci 49:S19–S23

Green P (2002) Documentation for phrap. Genome Center, University of Washington. http://www.phrap.org/

Hegyi H, Gerstein M (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 288:147–164

Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285:1887–1897

Mayer K, Murphy G, Tarchini R, Wambutt R, Volckaert G, Pohl P, Düsterhöft A, Stiekema W, Entian K-D, Terryn N, Lemcke K, Haase D, Hall CR, van Dodeweerd A-M, Tingey SV, Mewes H-W, Bevan MW, Bancroft I (2001) Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. Genome Res 11:1167–1174

Poethig RS (2001) Life with 25,000 genes. Genome Res 11:313–316

Rubin GM, Yandell MD, Wortman JR, Miklos GLG, Nelson CR, et al (2000) Comparative genomics of the eukaryotes. Science 287:2204–2215

Somerville C, Somerville S (1999) Plant functional genomics. Science 285:380–383

Wolf KH, Gouy M, Yang Y-W, Sharp PM, Li W-H (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205