Opinion TRENDS in Plant Science Vol.12 No.11

POFs: what we don't know can hurt us

Martin Gollery[1], Jeff Harper[2], John Cushman[2], Taliah Mittler[2] and Ron Mittler[2,3]

[1] TimeLogic – a Division of Active Motif, Incline Village, NV 89451, USA
[2] Department of Biochemistry and Molecular Biology, MS200, University of Nevada, Reno, NV 89557, USA
[3] Department of Plant Science, Hebrew University of Jerusalem, Givat Ram, Jerusalem 91904, Israel

Over a quarter of all eukaryotic genes encode proteins with obscure features that lack currently defined motifs or domains (POFs). Interestingly, most of the differences in gene repertoire among species were recently found to be attributed to POFs. A comparison of the *Arabidopsis*, rice and poplar genomes reveals that *Arabidopsis* contains 5069 POFs, of which 2045 have no obvious homologs in rice or poplar and are likely to be involved in species- or phylogenetic-specific functions in *Arabidopsis*. The study of POFs is an important endeavor that will shed much needed light on the genetic properties that make any given plant species unique. Furthermore, with respect to many species-specific features, such studies show that we seem to be limited in what we can expect to learn from a model plant such as *Arabidopsis*.

Why are POFs important?

The rapid proliferation of genomic and metagenomic sequencing data has drawn increasing attention to the large number of genes of unknown function, which seem to be an integral part of the genetic blueprint of most organisms [1–6]. On average, 15–40% of every eukaryotic genome sequenced to date contains genes that encode proteins with obscure features that lack currently defined motifs or domains (POFs; see Glossary) [2]. The recent metagenomic ocean-sequencing expedition uncovered tens of thousands of proteins with undefined features, highlighting how little we know of the amazing diversity of protein sequences [1]. But what are the roles of POFs? Is it possible that, buried within each new genome sequenced, there exists an entire set of pathways and genetic programs of which we are completely unaware? Here, we discuss the role of POFs in plants and highlight the possibility that they have a key role in determining ecologically and agronomically important species-specific features of different plants.

How to provide a measurable definition to the unknown?

Although the term 'gene of unknown function' is used broadly, it is difficult to define. How do we quantify the unknown in genes with unknown function?

Using overall amino acid sequences similarity

One way to define a 'gene of unknown function' is to use a similarity-based definition using a nucleotide or amino acid sequence comparison or equivalent. If a gene has no homolog in any other genome sequenced to date, or in any of the databases available [e.g. using a BLAST search against all sequences in the National Center for Biotechnology Information (NCBI) database], then it could be defined as an unknown. This type of classification has been used to define orphan open reading frames (ORFans) [3,5,6]. The total number of ORFans in the NCBI database was recently estimated to be ~80 000, underlying the magnitude of the problem that researchers face in understanding these genes and their roles [1]. However, at least two problems are associated with the ORFan definition: (i) although there are many proteins that have some degree of similarity among different genomes, this similarity fails to provide clues to a possible gene function; and (ii) the BLAST E-value cutoff used to classify ORFans is often not rigorously defined. Using too high (e.g. $>10^{-2}$), or too low (e.g. $<10^{-30}$) cutoff values could drastically change the annotation of an ORFan.

A similarity-based definition of an unknown can also be used to define a protein with a homolog(s) in other genomes or in available databases but these homologs have no known classification. Such genes are often referred to as expressed proteins with unknown function or hypothetical proteins. At least two problems are also associated with this type of definition: (i) the BLAST E-value cutoff used for the similarity search is often poorly defined; and (ii) the annotation of genes in the available databases is often inaccurate or outdated. The second problem is serious

Glossary

BLAST: a program that finds protein or nucleotide sequences that are similar to a target sequence. It provides two values: S and E. The S-score is a measure of the similarity between the query and the sequence. The E-value is a measure of the reliability of the S-score. The definition of the E-value is, therefore, the probability owing to chance that there is another alignment with a similarity greater than the given S-score.
Domain of unknown function (DUF): a domain that can be identified in a given protein by an HMM search but has no defined function.
Hidden Markov model (HMM): a type of probabilistic model used to align and analyze sequence datasets by generalization from a sequence profile; it is well suited to providing a mathematical framework for profile analysis.
Hypothetical protein: a predicted protein for which there is no experimental evidence that it is expressed *in vivo*.
Meta-genomic: study of the collective genomes of microorganisms (as opposed to clonal cultures). The technique is to sequence DNA obtained directly from the environment of the microorganism.
ORFan: orphan open reading frame (ORF) with no detectable sequence similarity to any other sequence in the databases.
Protein families (Pfam): a large collection of multiple sequence alignments and HMMs covering many common protein families.
Protein with defined feature (PDF): a protein that contains at least one previously defined domain or motif.
Protein with obscure features (POF): a protein that lacks currently defined motifs or domains.

Corresponding author: Mittler, R. (ronm@unr.edu).

 1360-1385/$ – see front matter doi:10.1016/j.tplants.2007.08.018

because many databases have missing annotations, lack manual annotations that are required to provide a quality check of the automated, computer-based annotations, or are in need of frequent and methodical updates. For example, updated information from recent studies, which provide genetic or biochemical clues to the function of a protein previously annotated as a hypothetical protein, is often not included in many of the available databases.

Using HMMPFAM

A different type of classification, which might help in the annotation of unknowns, is one that uses a hidden Markov model protein family (HMMPFAM) search to identify domains or motifs that can provide a putative function for a protein. For example, a protein can be defined as a putative protein kinase, based upon the presence of a protein kinase domain. A protein that lacks any previously defined domains or motifs can be classified as a POF. Although this definition is also not an ideal one to use to define an unknown, because functional information could be associated with a POF (e.g. it has a defined biochemical or genetic function, or it interacts with a protein of a known function), it has been used in several genome annotation projects [7].

Future definitions

It seems that there is currently a lack of an accurate or measurable definition of unknown proteins. Moreover, it could be argued that the broad use of this term among biologists is complicating the interpretations of future genome annotations, in addition to current and future experimental results. The answer to how to define unknown proteins is one that could be addressed with the availability of more sequencing data, the development of new and dynamic bioinformatics tools dedicated to this question and better annotation of databases. At present, we favor the use of POFs and ORFans, although these designations are moving targets that must be updated on a regular basis owing to the increasing influx of new sequencing data and the evolution of enhanced HMM models.

Why should we study POFs in plants?

The definition of POFs was recently used in a comparative study of ten different predicted proteomes, including *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Arabidopsis thaliana, Oryza sativa, Drosophila melanogaster, Anopheles gambiae, Caenorhabditis elegans, Mus musculus, Rattus norvegicus* and *Homo sapiens* [2]. Proteins were annotated as POFs, or proteins with defined features (PDFs), based on an HMMPFAM search (E-value threshold of 10^{-4}) against several major signature databases, including Pfam (http://www.sanger.ac.uk/Software/Pfam/), TIGRFAM (http://www.tigr.org/TIGRFAMs/), SMART (http://smart.embl-heidelberg.de/), and Superfamily (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY) (Figure 1a). A protein that had no match to any one of the models in any of the databases was classified as a POF, whereas a protein that had a match to one or more of the models in any one of the databases was classified as a PDF. Surprisingly, it was found that, on average, 60% of the POFs identified in the ten predicted proteomes were species specific, compared with only 7.5% of the PDFs. As a group, POFs seemed to be similar to PDFs in their relative contribution to biological functions, as indicated by their expression, participation in protein–protein interactions and association with mutant phenotypes [2]. However, POFs had more predicted disordered structures (i.e. regions that fail to show a consistent or defined structure in a crystallized protein [8–11]). Such regions show a strong correlation with biochemical studies that suggest their involvement in protein–protein interactions and also in providing key regions for regulating the activity of a protein through a structural conformation switch [8–11]. The high content of disordered structure in POFs supports a hypothesis that they function in species- or phylogenetically specific regulatory and signaling networks [2].

To evaluate the diversity among PDFs and POFs in rice and *Arabidopsis*, their sequence relatedness to each other was compared using BLAST (Figure 1a). The percentage of related proteins was plotted as a function of similarity cutoff thresholds that ranged from non-stringent (BLAST E-values $>10^{-6}$) to stringent (from 10^{-9} to 10^{-80} or less) (Figure 1b). This method of plotting similarity differences enables the visualization of reproducible differences between PDFs and POFs across a wide range of cutoff thresholds. As seen in Figure 1b, POFs consistently displayed a higher degree of divergence compared with PDFs between the two plant proteomes [2].

The recent publication of a third plant genome, poplar *Populus trichocarpa* (Pt) [12], prompted the testing of how the addition of this genome affected the previous findings with *Arabidopsis* (At) and rice (Os) [2]. As shown in Figure 1c, using the classification method described above (Figure 1a), >26% of the poplar predicted proteome comprised POFs, compared with 19% and 33% of the *Arabidopsis* and rice proteomes, respectively.

In a comparison among the three plant proteomes, ~75% of the unique proteins of poplar were POFs (using a BLAST E-value cutoff of 10^{-6}, an average standard cutoff used by many different researchers [1–3,7,13] (Figure 2a). Following the addition of poplar to the analysis, the number of *Arabidopsis*-unique POFs, which was previously found to be 3342 [2], decreased to 2045 (a decrease of 38%), whereas the number of rice-unique POFs, which was previously 19 031 [2], decreased to 15 210 (a decrease of 20%). At least with respect to unique POFs, the model dicot proteome (At) was, therefore, disproportionably affected compared with the model monocot proteome (Os), by the addition of sequence data for the poplar genome (Pt), which is a dicot tree species.

The disproportionably large number of POFs and unique POFs in rice, compared with *Arabidopsis* and poplar [2] (Figures 1c and 2a), prompted the testing of whether more POFs in rice exist as members of large protein families. Surprisingly, as shown in Figure 2b, most POFs in all plant proteomes tested seemed to be encoded by single-copy genes and did not belong to members of large protein families. Taking into account the higher proportion of disordered structures in POFs, the shorter length of POFs compared with PDFs, and the low level of sequence similarity between POFs from different organisms, it was previously proposed that POFs could represent newly

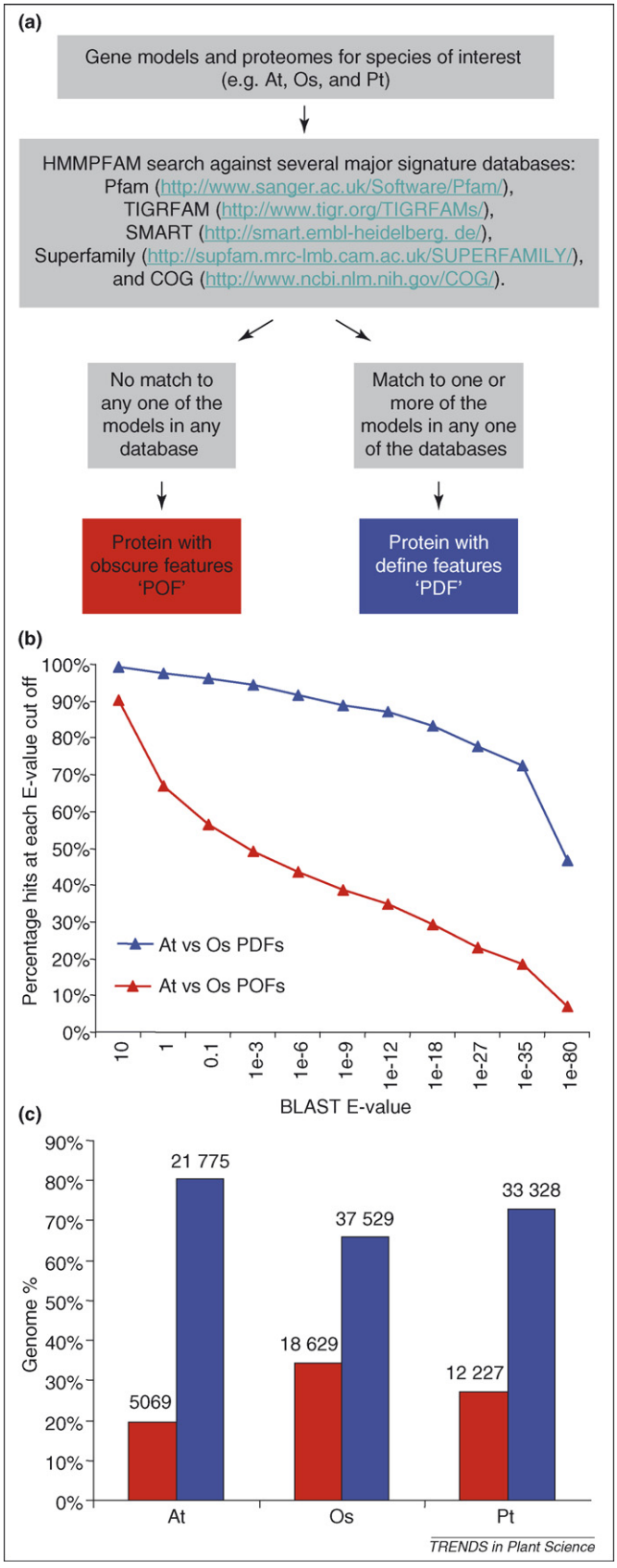

Figure 1. Identification of POFs in the predicted proteomes of species of interest, including At, Os and Pt. **(a)** The classification of PDFs and POFs in the three plant proteomes. **(b)** Relative similarity between POFs and PDFs from At and Os. The percentage of related proteins (i.e. proteins that find at least one hit) was plotted as a function of similarity cutoff thresholds that ranged from non-stringent (BLAST E-values $>10^{-6}$) to stringent (from 10^{-9} to 10^{-80} or less). **(c)** Representation of POFs (red bars) and PDFs (blue bars) in the three plant proteomes. The total number of proteins in each group is given on top of each bar.

evolving genes or genes that are evolving much faster than the genome average [2]. The finding that most POFs exist in plants as singletons (Figure 2b) seems to support this hypothesis. What, then, is the difference between poplar, *Arabidopsis* and rice? Why would rice have a higher number of POFs? At least two different factors could differentiate between rice and the two other plant proteomes: (i) the annotation of genes in rice could be incomplete [13]; and (ii) in contrast to *Arabidopsis* and poplar, rice has been the subject of almost 4000 years of selection and breeding [14]. Could either breeding or inaccurate annotation be the cause of the high number of POFs in rice? Further studies are required to address these possibilities.

POFs seem to account for most of the species-specific differences between the three plant proteomes. As shown in Figure 3, POFs, which account for only 19–33% of the total number of proteins in each proteome (Figure 1c), have, on average, a threefold higher representation in the class of unique proteins for each organism (Figure 2a). The analysis shown in Figure 3 reveals that the three plant proteomes have >1800 POF groups in common. In contrast to the POFs that are unique to each proteome, these shared POFs are likely to have generalized functions, such as those involved in cellular structure and general metabolism or defense common among most plants. Moreover, it is expected that domains of unknown function (DUFs) will eventually be generated for most of the shared POFs and deposited in major signature databases, such as Pfam, resulting in an improved annotation of POFs and PDFs, and the shrinking of the POFs pool [2].

What more can be done to study POFs in plants?

Several different avenues of research might accelerate the annotation of POFs in plants, including further sequencing of plant genomes, coupled with improved annotation of databases, large-scale mutant and transgenic characterization for different phenotypes, such as growth, development and tolerance to environmental stresses, and the generation of comprehensive databases for the tertiary structure of proteins that would add a new dimensional tool for comparing among different genomes (i.e. structural genomics) [1–6]. An additional reason why proteins such as POFs are currently undefined might be related to the sociology of science. In general, scientists have focused their studies on relatively few organisms, and devoted most of their resources to indepth analysis of relatively few proteins. These are often chosen because of their general relevance to fundamental questions in a broad group of organisms or because they exhibit strong evolutionary conservation. By contrast, the study of a species-specific protein is often a lonely pursuit. Another source of bias could lie in the tendency, inherent to classical biochemical methods, to be strongly biased towards the characterization of folded, active proteins that have highly ordered structures (e.g. PDFs) and for which structural information is more readily obtained. By contrast, disordered proteins (e.g. POFs) are less well studied because

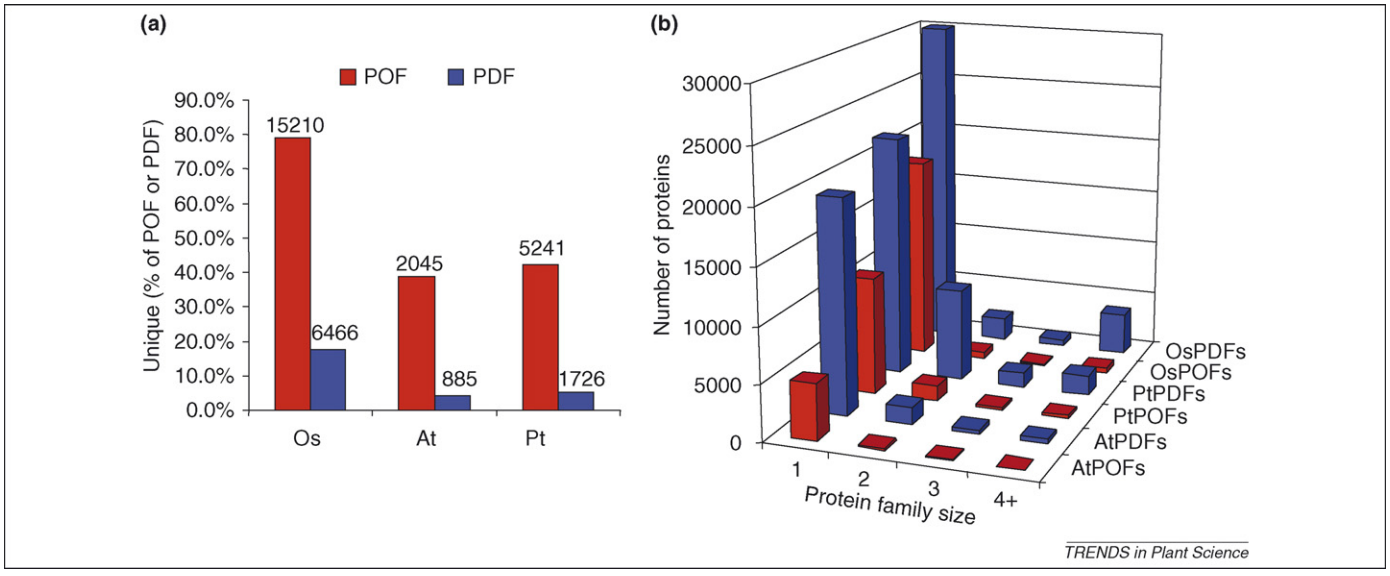

Figure 2. POFs account for most unique proteins among the At, Os and Pt proteomes, and are mainly encoded by single-copy genes. **(a)** Distribution of POFs (red bars) and PDFs (blue bars) among proteins unique to each of the three different plant proteomes. A BLAST cutoff E-value of 10^{-6} was used to define a protein as unique (see Gollery *et al.* [2] for more details). **(b)** POFs and PDFs in protein families from the three different plant proteomes. A BLAST E-value cutoff of 10^{-6} was used to assign a protein to a particular protein family [2].

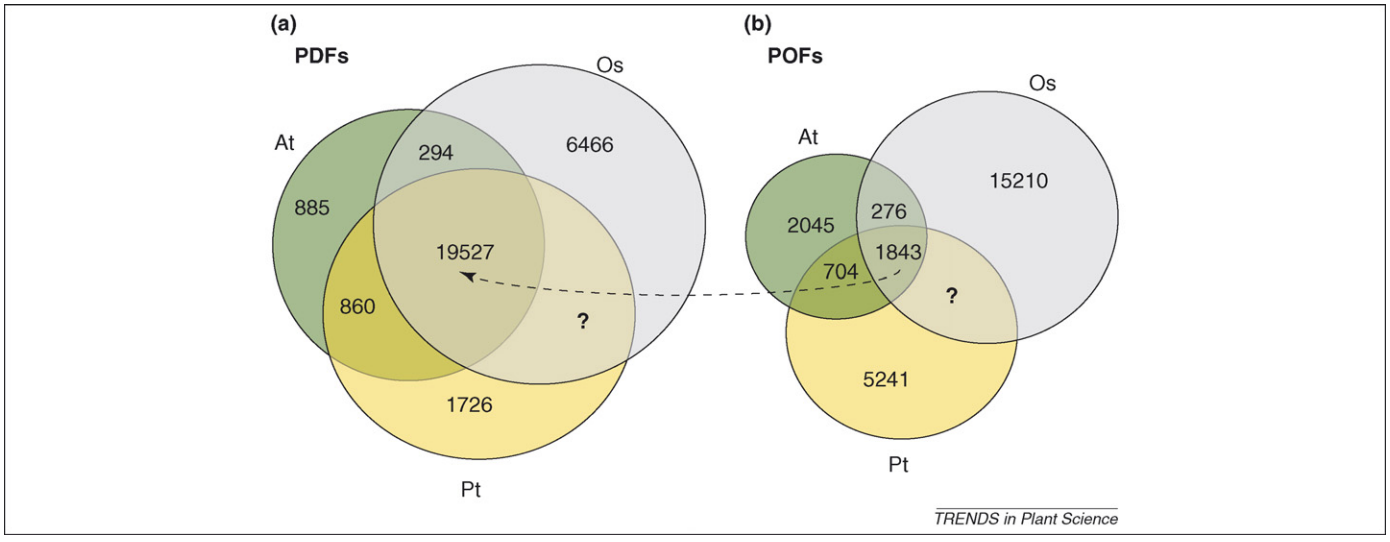

Figure 3. The distribution of **(a)** PDFs and **(b)** POFs among proteins common or unique to the At, Os and Pt proteomes (not to scale). Because the At predicted proteome is used as the reference point, and because it is the best annotated plant proteome to date, the numbers of proteins that are common to Os and Pt, but do not appear in At, could not be accurately predicted and are, therefore, indicated by '?'. The dashed arrow indicates the common POFs that could be annotated with new DUFs and eventually be classified as PDFs [2].

they lack a readily recognized activity, and structural information is more difficult to obtain for them [2].

Conclusion

Understanding the origins of species specificity was recently ranked among the top 25 scientific questions of our time, by the journal *Science* [15]. The increased availability of genome sequences has re-energized this effort and enabled large-scale genome comparative studies. The findings discussed here support a general expectation that understanding the unique biology of a given plant species will ultimately involve understanding the functions of an unexpectedly large number of proteins that: (i) have no defined motifs or domains; (ii) are likely to have significant regions of disordered structure; and (iii) are restricted to a single species or a closely related phylogenetic branch [2].

Although more challenging than studying conserved or defined proteins (e.g. PDFs), we believe that the study of POFs is an important endeavor that will shed much needed light on the specific properties that makes a given plant species unique. Furthermore, with respect to many species-specific features, we seem to be limited in what we can expect to learn from a model plant such as *Arabidopsis*. Thus, for example, studying an *Arabidopsis* unique POF might not contribute to our understanding of what makes a rice plant develop in the specific manner that it does, and vice versa. POFs could also account for ecologically important attributes that are specific for different plants. Understanding many ecologically or agronomically important species-specific aspects of plant biology, therefore, will require unraveling the function of many species-specific POFs in different plants.

Acknowledgements
Supported by the National Science Foundation (NSF-0420033; NSF-0420152). This publication was made possible by NIH Grant Number P20 RR-016464 from the IDeA Network of Biomedical Research Excellence (INBRE) Program of the National Center for Research Resources (NCRR) through its support of the Nevada Center for Bioinformatics.

References

1 Yooseph, S. *et al.* (2007) The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16
2 Gollery, M. *et al.* (2006) What makes species unique? The contribution of proteins with obscure features. *Genome Biol.* 7, R57
3 Siew, N. and Fischer, D. (2004) Structural biology sheds light on the puzzle of genomic ORFans. *J. Mol. Biol.* 342, 369–373
4 Roberts, R.J. (2004) Identifying protein function – a call for community action. *PLoS Biol.* 2, 293
5 Galperin, M.Y. and Koonin, E.V. (2004) 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32, 5452–5463
6 Fischer, D. and Eisenberg, D. (1999) Finding families for genomic ORFans. *Bioinformatics* 15, 759–762
7 Chothia, C. *et al.* (2003) Evolution of the protein repertoire. *Science* 300, 1701–1703
8 Dunker, A.K. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59
9 Dunker, A.K. *et al.* (2002) Intrinsic disorder and protein function. *Biochemistry* 41, 6573–6582
10 Gunasekaran, K. *et al.* (2003) Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* 28, 81–85
11 Brown, C.J. *et al.* (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* 55, 104–110
12 Tuskan, G.A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604
13 Sterck, L. *et al.* (2007) How many genes are there in plants (. . . and why are they there)? *Curr. Opin. Plant Biol.* 10, 199–203
14 Doebley, J.F. *et al.* (2006) The molecular genetics of crop domestication. *Cell* 127, 1309–1321
15 Pennisi, E. (2005) What determines species diversity? *Science* 309, 90