Bioinformatics as a critical prerequisite to transcriptome and proteome studies

Elisabeth Jamet*

Surfaces Cellulaires et Signalisation chez les Végétaux, UMR CNRS-UPS 5546, 24, chemin de Borderouge, BP 17 Auzeville, F-31326 Castanet Tolosan, France

Received 24 February 2004; Accepted 2 June 2004

Abstract

Large-scale genomic studies rely strongly on annotations available in databases to design experimental supports such as arrays or to explain results in term of biological meaning. Most of this information originates from bioinformatic predictions. Their accuracy as well as their relevance to existing biological data are critical in avoiding the misinterpretation of experimental results.

Key words: Bioinformatic predictions, biological data, database, genomic studies, interpretation of results.

Introduction

The increasing number of sequences that are available in databases, a hundred times higher than ten years ago (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html), makes the accuracy of sequence annotation a great challenge. By contrast with global analyses of transcriptional activity that aim to scan the genome for potential transcription units (Choudhary et al., 2001; Yamada et al., 2003), transcriptome and proteome studies require the structure and function of genes to be determined precisely. Transcriptome studies need arrays designed to follow the expression of specific collections of genes that must be relevant to the biological question addressed. Proteomic approaches rely on the identification of proteins performed using mass spectrometry either from peptide sequencing or from peptide mass fingerprinting.


This paper will provide some examples of misleading annotations with regard to putative protein function that may cause mistakes either in array design or in data interpretation. Examples will be taken mainly from A. thaliana and from published papers or databases such as Uniprot, NCBI, TAIR, TIGR, and MIPS.

Proteins rich in particular amino acids

Cell wall structural proteins provide interesting examples of poor quality annotation because their sequences are rich in particular amino acids. Three classes of structural proteins have been clearly defined: extensins characterized by the presence of numerous Ser-Pro$_n$ $(n\geq3)$ motifs separated by Tyr-, Lys-, His, and Val-rich regions (Kieliszewski and Lamport, 1994); Hydroxyproline/Proline-Rich proteins (H/PRPs) characterized by a high content in Pro and Pro-Pro-X-Y-Lys motifs, where X, Y=Val, Tyr, His, or Glu (Showalter, 1993); and Glycine-Rich proteins (GRPs) characterized by a high content in Gly (up to 70%) organized in repeats of the (Gly-X) motif, where X=Gly,

* Fax: +33 (0)5 62 19 35 02. E-mail: jamet@scsv.ups-tlse.fr
Ala, or Ser (Showalter, 1993). Numerous proteins predicted to have a signal peptide by PSORT (http://psort.nibb.ac.jp/form.html) and TargetP (http://www.cbs.dtu.dk/services/TargetP) and showing only short stretches of Pro or Gly have been wrongly annotated as extensin-like, PRP or GRP. This is notably the case for At2g33790 (14.6% Pro), At5g26070 (23.5% Pro), and At4g28300 (13.6% Pro) annotated as extensins or PRPs in the Uniprot, NCBI, TAIR, and TIGR databases. At4g34300 (14.7% Gly), At5g26070 (14.6% Gly), and At2g15340 (17.6% Gly) are presently annotated as GRPs in the NCBI, TAIR, and TIGR databases, but as putative or unknown proteins in the Uniprot and MIPs databases. Other examples are provided by a recent transcriptome study on peach by Trainotti et al. (2003). Contig 010 shows homology to the S65062 cotton fiber protein 6 (John, 1996). Since, this protein has only one short Ser-Gly motif, it cannot be classified among structural proteins as suggested by the authors. In the same way, contig 125 shows homology to Arabidopsis thaliana NP_176440 (At1g62510). The primary sequence of the encoded protein has only one short X-Pro (with X = His, Lys, Asn, Thr, Ser) domain that again is not sufficient to classify it among the structural proteins mentioned in the MIPs database. It actually comprises a PFAM domain (PF00234) defining a protease inhibitor/seed storage/LTP family (http://hits.isb-sib.ch/cgi-bin/PFSCAN) clearly indicated in the NCBI, TAIR, and TIGR databases.

Proteins having several biological activities

An example is provided by a protein family encoding putative Asp proteases. It comprises about 30 members sharing the IPR009007 domain for peptidase aspartic (http://www.ebi.ac.uk/InterProScan/). Most of them are predicted to be secretory proteins by PSORT and TargetP. Four of them (At1g09750, At5g07030, At3g54400, and At3g61820) were identified in proteomic studies on the A. thaliana primary cell wall (Borderies et al., 2003; G Boudart et al., unpublished results). Most of them are currently annotated as chloroplast nucleoid or nucleoid DNA binding-related proteins in the NCBI and MIPS databases. These proteins are actually homologous to a tobacco chloroplast nucleoid DNA-binding protein that was shown to have a protease activity (Murakami et al., 2000). The annotation of A. thaliana proteins in the NCBI and MIPS databases thus appears to be misleading. By contrast, all these proteins are correctly annotated as Asp proteases in Uniprot as well as in the A. thaliana TIGR and TAIR databases.

Proteins containing several functional domains

Proteins containing several functional domains may be a problem when results of sequence comparison or functional domain search are not carefully interpreted. A first example is the protein encoded by At3g22060. It is presently annotated as a receptor protein kinase related in the NCBI, TAIR, and TIGR databases because it contains a PFAM profile named Domain of Unknown Function DUF 26 (PF01657) usually associated with the protein kinase domain PFAM (PF00069) not present in this protein. The protein has therefore no predictable function at the moment. A second example is that of proteins belonging to a family of curculin-like (mannose-binding) lectins (At1g78850, At1g78860, and At1g16900). It is mentioned in the NCBI, TAIR, TIGR, and MIPS databases that they show low similarity to a Ser/Thr protein kinase of Zea mays (GI: 2598067). This similarity does exist with the curculin-like (mannose-binding) lectin domain (PF01453), but not with the protein kinase domain (PF00069) of the Z. mays protein absent in many members of the lectin family. The same is true for At1g53070 that is annotated as protein kinase related. The encoded protein has a legume lectin beta domain (PF00139) and no protein kinase domain. The annotation of genes At1g78850 and At1g53070 misled the authors of a proteomic study discussing the presence of putative protein kinases in cell wall preparations (Chivasa et al., 2002; Ndimba et al., 2003). They actually found putative lectins with completely different biological functions.

Conclusion

All the above-mentioned misleading annotations originate from the misinterpretation of sequence comparisons or domain searches. A careful and critical bioinformatic analysis of DNA and/or protein sequences therefore appears to be an absolute requirement before starting a transcriptome analysis or discussing the results from a proteomic analysis. The importance of comparing results obtained with different bioinformatic softwares has been clearly shown in the Aramemnon database which was especially designed to collect integral membrane proteins (http://aramemnon.botanik.uni-koeln.de/) (Schwacke et al., 2003). It integrates data from 11 trans-membrane predictions and 8 signal peptide predictions and illustrates the type of discrepancies that may be observed between the results. Moreover, the relevance of bioinformatic predictions to biological data should be checked whenever possible to prevent mistakes. Biological data have been proved to be essential for the improvement of the quality of genome annotations, as recently shown by the systematic sequencing of full-length cDNAs (Haas et al., 2002), the use of oligonucleotide tilling arrays (Yamada et al., 2003), and proteomics (Choudhary et al., 2001; Borderies et al., 2003).

Acknowledgements

The author is grateful to Professor Rafael Pont-Lezica for critical reading of the manuscript. The laboratory is supported by the Centre...
National de la Recherche Scientifique and the Université Paul Sabatier, Toulouse, France.

References


