Bioinformatics as a critical prerequisite to transcriptome and proteome studies.

Elisabeth Jamet

Published in Journal of Experimental Botany 55 (2004) 1977-1979.

PMID: 15258172

Bioinformatics as a critical prerequisite to transcriptome and proteome studies

Elisabeth JAMET

Elisabeth JAMET
Surfaces Cellulaires et Signalisation chez les Végétaux
UMR CNRS-UPS 5546
24, chemin de Borderouge
BP 17 AUZEVILLE
31326 CASTANET TOLOSAN
Tel: +33 (0)5 62 19 35 30
Fax: +33 (0)5 62 19 35 02
e-mail : jamet@scsv.ups-tlse.fr

Abstract

Large-scale genomic studies strongly rely on annotations available in databases to design experimental supports such as arrays or to explain results in term of biological meaning. Most of these informations originate from bioinformatic predictions. Their accuracy as well as their relevance to existing biological data become critical to avoid misinterpretation of experimental results.

Introduction

The increasing amount of sequences 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. In contrast to global analyses of transcriptional activity that aim at scanning the genome for potential transcription units (Choudhary et al., 2001; Yamada et al., 2003), transcriptome and proteome studies require structure and function of genes to be precisely determined. Transcriptome studies require 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 *via* mass spectrometry either from peptide sequencing or peptide mass fingerprinting.

Bioinformatics analyses are now made easier since the quality of the available softwares and of the annotations provided by databases is continuously improving, especially for plant model organisms like *Arabidopsis thaliana* and *Oryza sativa*. Informations related to gene structure and expression are available at NCBI (http://www.ncbi.nlm.nih.gov/Database/index.html). Genomic, cDNA and EST sequences are compared to establish the exon/intron structure of genes. Bioinformatic predictions are checked and eventually corrected using primary sequences (http://www.ncbi.nlm.nih.gov/RefSeq/) (Pruitt and Maglott, 2001). Informations on protein functions are provided in several databases including Uniprot where the origin of the proposed function is mentioned (experimental or electronic annotation) (http://www.expasy.uniprot.org/) (Apweiler *et al.*, 2004), NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide), TAIR

(http://www.arabidopsis.org/), TIGR (http://www.tigr.org/tdb/e2k1/ath1/) and MIPS (http://www.mips.biochem.mpg.de/proj/thal/db/index.html).

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 mainly taken 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 due to their sequences that 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 ($n \geq 3$) 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, 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 streches 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), At4g33930 (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 is again not sufficient to classify it among structural proteins as mentioned in the MIPS database. Actually, it 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 thirty 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 *A. thaliana* primary cell wall (Borderies *et al.*, 2003; Boudart *et al.*, unpublished results). Most of them are presently annotated as chloroplast nucleoid or nucleoid DNA binding-related proteins in the NCBI and MIPS databases. Actually, these proteins are homolog 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. On the contrary, 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 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. Same remark 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 mentioned misleading annotations originate from 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 results from a proteomic analysis. The importance of comparing results obtained with different bioinformatic softwares is well shown in the Aramemnon database 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. Indeed, biological data proved to be essential to improve the quality of genome annotations as recently shown by systematic sequencing of full-length cDNAs (Haas *et al.*, 2002), 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 Prof 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

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. 2004. UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Research* **32**, D115-D119.

Borderies G, Jamet E, Lafitte C, Rossignol M, Jauneau A, Boudart G, Monsarrat B, Esquerre-Tugaye MT, Boudet A, Pont-Lezica R. 2003. Proteomics of loosely bound cell wall proteins of Arabidopsis thaliana cell suspension cultures: a critical analysis. *Electrophoresis* **24**, 3421-32.

Chivasa S, Ndimba BK, Simon WJ, Robertson D, Yu XL, Knox JP, Bolwell P, Slabas AR. 2002. Proteomic analysis of the *Arabidopsis thaliana* cell wall. *Electrophoresis* **23**, 1754-1765.

Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS. 2001. Matching peptide mass spectra to EST and genomic DNA databases. *TRENDS in Biotechnology* **19**, S17-S22.

John ME. 1996. Structural characterization of genes corresponding to cotton fiber mRNA, E6: reduced E6 protein in transgenic plants by antisense gene. *Plant Molecular Biology* **30,** 297-306.

Haas BJ, Volfovsky N, Town CD, Troukham M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL. 2002. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biology* **3**, 0029.1-0029.12.

Kieliszewski MJ, Lamport DT. 1994. Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. *The Plant Journal* **5,** 157-72.

Murakami S, Kondo Y, Nakano T, Sato F. 2000. Protease activity of CND41, a chloroplast nucleoid DNA-binding protein, isolated from cultured tobacco cells. *FEBS Letters* **468**, 15-18.

Ndimba BK, Chivasa S, Hamilton JM, Simon WJ, Slabas AR. 2003. Proteomic changes in the extracellular matrix of Arabidopsis cell suspension cultures induced by fungal elicitors. *Proteomics* **3**, 1047-1059.

Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* **29**, 137-140.

Schwacke R, Schneider A, van der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flugge UI, Kunze R. 2003. ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiology* **131**, 16-26.

Showalter AM. 1993. Structure and function of plant cell wall proteins. *The Plant Cell* **5,** 9-23.

Trainotti L, Zanin D, Casadoro G. 2003. A cell wall-oriented genomic approach reveals a new and unexpected complexity of the softening in peaches. *Journal of Experimental Botany* **54**, 1821-1832.

Yamada et al. 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842-846.