TECHNICAL BRIEF

Simple proteomic checks for detecting noncoding RNA

Stéphane Cruveiller[1], *Oliver Clay*[2], *Kamel Jabbari*[3] *and Giorgio Bernardi*[2]

[1] Atelier de Génomique Comparative, Genoscope, Centre National de Séquençage, Evry, France
[2] Stazione Zoologica Anton Dohrn, Naples, Italy
[3] Département de Biologie, CNRS FRE 2910, Ecole Normale Supérieure, Paris, France

Proper validation can accelerate sequence-based discovery of proteins and protein-coding genes. Databases currently contain a backlog of experimentally unverified gene models and tentative assignments of observed transcripts to coding or noncoding RNA. We present and apply a general principle, founded on base composition and the genetic code and validated here by bulk 2-D gels, that can improve the reliability of such classifications and of the algorithms or pipelines that lead to them.

Received: October 24,2006
Accepted: October 31, 2006

Keywords:
2-DE / Base composition / Isoelectric point / Protein predictions

In mammals, noncoding transcript species are apparently at least as common as coding ones [1], and a "staggering" 62% of the mouse genome is transcribed [2]. This recent discovery has a new implication for gene prediction pipelines: the mere presence of a proper transcript gives only very limited support to the hypothesis of a protein product. The same conclusion can be rephrased positively: there exist many more candidates for possibly functional noncoding RNAs than were previously assumed.

When correctly used, base composition of DNA/RNA or, correspondingly, conceptual amino acid usage can help to assess if transcripts are likely or unlikely protein precursors. This principle is often overlooked, despite much attention to automating gene prediction. In previous work, we proposed a simple screen, and applied it to rice [3, 4], based on a constraint that is experienced by protein-coding genes in species ranging from human to *Escherichia coli*. GC levels (guanine + cytosine %) are distinctly lower in second (GC2) than in third (GC3) codon positions. RNA transcribed from noncoding DNA has, however, no codon positions and therefore no contrasts between them. As a consequence, 2-D scatterplots show real protein-coding genes clustering along an evolutionarily conserved line that departs strongly from the main diagonal GC2 = GC3, where incorrectly predicted genes line up with noncoding DNA or RNA. This principle was applied to show that the new genes predicted for rice, which had no homologs in other sequenced species and amounted to half the putative gene set, were unlikely to code for proteins that function *in vivo* [3, 5]. Recent rice gene counts [6, 7] indeed no longer include most of those *de novo* predictions.

The check at the DNA/RNA level can be translated to the proteomics level in order to allow bulk experimental verification of putative protein sets. In species that use the standard genetic code, GC2 is just the summed frequencies of eight amino acids: Arg, Ser, Thr, Pro, Ala, Gly, Cys, and Trp. Proteins encoded by unusually high GC2 would therefore show up also in 2-D proteomics gels, since high GC2 leads to a high p*I*, (pH): only arginine represents the GC2 class in a basic/acidic quotient, (Lys + Arg)/(Glu + Asp), that dominates the p*I* [8, 9] (see Supplementary Material for details).

Methods are available for simulating real 2-D gels *in silico*. Typically such theoretical or virtual 2-D gels [10–13] are similar to each other and to those obtained by using simple equations [8, 9], but certain corrections can temper the unrealistically abrupt bimodality of p*I*s that the equations would predict.

We used the program JVirGel [10] to check how easily sets of predicted coding regions on transcripts would deviate, on a 2-D gel, from known, *bona fide* protein-coding RNA. Figure 1 shows the difference between expected positions of

Correspondence: Professor Giorgio Bernardi, Stazione Zoologica Anton Dohrn, 80121 Naples, Italy
E-mail: bernardi@szn.it
Fax: +39-081-764-1355

Abbreviations: GC2, guanine + cytosine % in second codon positions of protein-coding genes; **GC3,** guanine + cytosine % in third codon positions of protein-coding genes

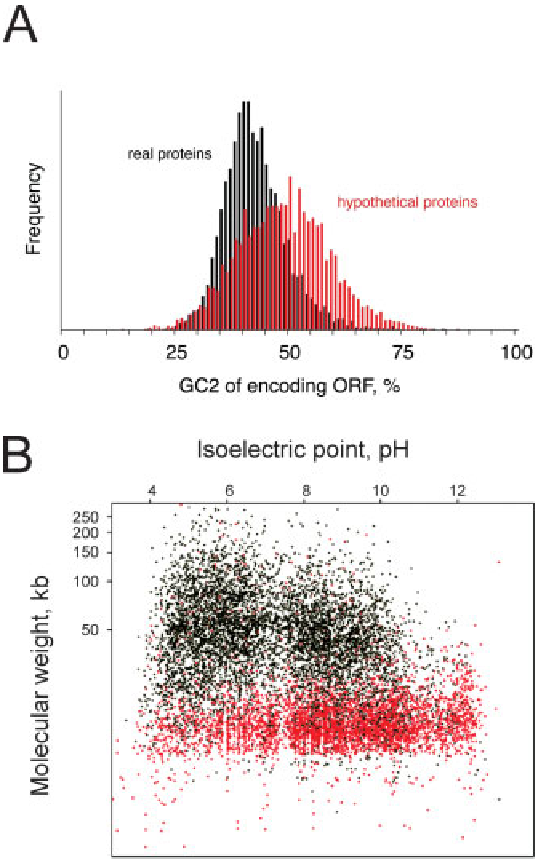

Figure 1. Large compositional differences between known (black) and purely hypothetical (red) human proteins that are supported by transcribed RNAs, as seen in GC2 distributions (top, A) and *in silico* 2-D gels (bottom, B). GC2 can be calculated from coding DNA/RNA sequences or, alternatively, from the protein sequences they encode. Human-curated sequences from the H-invitational database [14] were processed by JVirGel [10] using default settings. The known protein set ($n = 6207$) contained no "questionable", hypothetical, only domain-containing, "similar to known", or "family" proteins. The purely "hypothetical" protein set ($n = 5027$) contained no "conserved hypothetical" sequences.

purely hypothetical *versus* known human proteins from the H-invitational transcript database [14], chosen here because it classifies transcripts according to the evidence for their coding status. The purely hypothetical proteins concentrate thickly in a quite small region (lower right) corresponding to very short, and often very basic (high p*I*) proteins. By contrast, entire, real proteins are very rare in this region and are dispersed elsewhere.

The accumulated data now available from experimental proteomics all point to essentially the same situation in real gels also. Unlike the hypothetical proteins shown in Fig. 1, typical databases of human 2-D gels [15, 16] show very few if any spots in the most basic sector, and instead a high proportion of spots in the more acidic regions of the gel. In fact, there has been a long hunt for the most basic proteins (pH 10–13) that genome sequencing projects predicted. This hunt remained unsuccessful despite the targeted improvements in electrophoretic methods that it motivated [12, 17]. Especially with such improvements, one would have expected a reasonable portion of the "missing proteins" to become visible as spots, patches, or smears.

The parsimonious conclusion, from the above considerations, is that the sizeable numbers of predicted proteins with high to very high p*I* are not just perhaps hard to detect, but that they are absent. In other words, although *ab initio* predicted coding sequences in observed transcripts are certainly a useful resource for gene and protein discovery [18], a large proportion of the hypothetical proteins shown in Fig. 1 (red), and of many other predicted proteins like them, are unlikely to exist *in vivo* except possibly in exceedingly low amounts. Indeed, invisibility on a standard 2-D gel implies, at best, a presence *per* protein species that is far below the femtomolar quantities needed for MS [19]. Such an acknowledgement opens the door for more promising analyses of large quantities of already sequenced transcripts, under a working assumption of possibly functional, noncoding RNA. In this context, simple plots at the DNA/RNA [3, 5] and/or protein level, such as those shown here, should facilitate the prediction of RNAs' coding or noncoding status (see Supplementary Material for more details). Finally, fine-tuning the links between proteomic experiments and the compositional statistics of DNA/RNA should allow answers to open evolutionary questions, such as whether or when 2-D gels of proteomes can be used for phylogeny [13, 20], and also to open practical questions, such as the number of protein-coding genes that contribute to the human proteome [21].

References

[1] Carninci, P., Kasukawa, T., Katayama, S., Gough, J. *et. al.*, *Science* 2005, *309*, 1559–1563.

[2] Claverie, J. M., *Science* 2005, *309*, 1529–1530.

[3] Jabbari, K., Cruveiller, S., Clay, O., Le Saux, J., Bernardi, G., *Trends Plant Sci.* 2004, *9*, 281–285.

[4] Cruveiller, S., Jabbari, K., Clay, O., Bernardi, G., *Genome Res.* 2004, *14*, 886–892.

[5] Cruveiller, S., Jabbari, K., Clay, O., Bernardi, G., *Brief. Bioinform.* 2003, *4*, 43–52.

[6] Bennetzen, J. L., Coleman, C., Liu, R., Ma, J., Ramakrishna, W., *Curr. Opin. Plant Biol.* 2004, *7*, 732–736.

[7] International Rice Genome Sequencing Project, *Nature* 2004, *436*, 793–800.

[8] Patrickios, C. S., *J. Colloid Interface Sci.* 1995, *175*, 256–260.

[9] Patrickios, C. S., Yamasaki, E. N., *Anal. Biochem.* 1995, *231*, 82–91.

[10] Hiller, K., Schobert, M., Hundertmark, C., Jahn, D., Münch, R., *Nucleic Acids Res.* 2003, *31*, 3862–3865.

[11] Gasteiger, E. *et al.*, in: Walker, E. M. (Ed.), *The Proteomics Protocols Handbook*, Humana Press, USA 2005.

[12] Bae, S.-H. *et al.*, *Proteomics* 2003, *3*, 569–579.

[13] Knight, C. G., Kassen, R., Hebestreit, H., Rainey, P. B., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 8390–8395.

[14] Imanishi, T. *et al.*, *PLoS Biol.* 2004, *2*, 0856–0875.

[15] Thiede, B., Siejak, F., Dimmler, C., Jungblut, P. R., Rudel, T., *Electrophoresis* 2000, *21*, 2713–2720.

[16] Finehout, E. J., Franck, Z., Lee, K. H., *Electrophoresis* 2004, *25*, 2564–2575.

[17] Görg, A., Obermaier, C., Boguth, G., Weiss, W., *Electrophoresis* 1999, *20*, 712–717.

[18] Yamasaki, C. *et al.*, *Gene* 2005, *364*, 99–107.

[19] Gygi, S. P. *et al.*, *Proc. Natl. Acad. Sci. USA* 2000, *97*, 9390–9395.

[20] Brocchieri, L., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 8257–8258.

[21] Southan, C., *Proteomics* 2004, *4*, 1712–1726.

Errata

Simple proteomic checks for detecting noncoding RNA

By O. Clay *et al.*, vol. 7 issue 3, pp. 361–363
DOI: 10.1002/pmic.200600813

In Figure 1 of this paper, the vertical axis shows molecular weights in kDa, not kb.

Cover Illustration of Issue 4, vol. 7

Please note that the cover illustration of the torpedo is provided by courtesy of Mr Phillip Colla, Natural History Photography, Carlsbad, CA, USA.