Correspondence

Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data?

Adam Pavlíček[a], Oliver Clay[b], Giorgio Bernardi[b,*]

First published online 27 June 2002

The contribution of transposable elements (TEs), including Alus, to human coding sequences has recently been reported to be high, 4% (1.3% Alus) out of 13 799 sequences [1,2]. This is surprising, because previous examinations had revealed only very few repeats, and almost no Alus, in coding sequences [3,4,25]. Since extreme caution about input data has been suggested [5–7], we examined the database of [1] and found that many (~30%) of its TE-containing sequences or their protein products are defined as 'hypothetical', and 63% (421/669 sequences) are annotated as 'predicted, without experimental evidence or records without final NCBI revision'. Such a dataset is likely to contain several sequences that remain untranscribed, and more that remain untranslated. Not even experimental validation [8], let alone computer prediction of functional genes is foolproof: the errors in coding sequence databases such as those used in [1] may well amount to 1–2% or more.

Essentially all reported coding regions derived from Alus, or containing alternatively spliced Alus, have been detected at the RNA (cDNA) level, instead of at the protein level [3,9]. In eukaryotic cells, there is a significant turnover of RNA, and several steps of quality control exist for the synthesized RNA in both nucleus and cytoplasm [10–14]. mRNAs with an aberrant 3′ end are generally retained and/or degraded at their site of transcription [15] and the majority of stable RNA polymerase II transcripts remain in the nucleus as 'junk' RNA, so they never reach the cytoplasm [10]. The minority of transcripts that are successfully exported from the nucleus undergo additional check(s) during their translation. For example, there are specialized degradation mechanisms for transcripts having premature stop codons or lacking terminal codons, which prevent the creation of aberrant, potentially pathogenic proteins [11,13,16]. Thus, even detection of a transcript at the mRNA (cDNA) level cannot guarantee that these mRNAs are ever translated into stable proteins. As has been summarized in the light of growing evidence [17], 'mRNA abundance is a poor indicator of the levels of the corresponding protein', yet 'it is the proteome that determines cell phenotype': the transcriptome does not faithfully represent the proteome. Furthermore, to become a viable protein, a transcript must (after its accurate translation and possible post-translational modification) resist degradation until it can serve its functional role at the site of its required action. These facts underline the importance of detection at the protein level, for elucidating whether SINEs or other repeats contribute to true coding sequences in humans or mice.

The most accurate sources of proteins are 3D structure databases and direct amino acid sequencing. Out of 781 non-redundant human proteins from a 3D database or determined at the amino acid level that we extracted from [18] (mean length 404 aa; including some fragments, but neglecting all peptides shorter than 50 aa or having >70% identity) and compared to human repeats in RepBase [19] using TFASTX [20], we found no Alu-related protein domain (the best hit has an *E*-value of 0.5). Twenty-eight apparently significant hits with *E*-values under 0.01 were detected, but mainly from protein-coding elements (DNA transposons and LINE1). When cDNAs encoding these 28 proteins were extracted and searched by RepeatMasker [21], no interspersed repeats were detected. In addition, the similarity regions that had been reported by TFASTX were also found in other vertebrate orthologs. In summary, we did not detect any repeat sequence in our dataset of 781 protein sequences.

In 1994, it was pointed out [5] that a discovery of a translated Alu element(s) in a functional part of a functional human protein 'would represent the first report of its kind and would have important evolutionary implications'. Despite the 7 years since this challenge, confirmed cases of Alu-containing sequences that encode a functional protein still remain extremely elusive.

The paucity of documented examples is a good indication that proteins are unlikely to utilize domains encoded by Alus for functional ends. The reluctance to accept this view is understandable, given the huge proportion of interspersed repeats in the human genome (around 45% [4]): in principle, at least some of them might have been recruited for functional purposes at the protein level. The great majority of previously detected repeat-derived coding sequences comes, however, from protein-coding repeats, and particularly from DNA transposons [4,25]. LINEs are less common in coding sequences and only a few Alus had been identified prior to the analysis of [1,2]. Since SINEs are derived from RNA genes without protein-coding capacity, the lack of Alu-encoded proteins is consistent with the notion that new domains arise from existing sequences encoding functional proteins (for example, by exon shuffling) and that the de novo creation of coding sequences from non-coding DNA is rare. Indeed, in the words of Graur and Li [22], 'True novelty is almost unheard of during evolution; rather, preexisting genes and parts of genes [presumably encoding functional proteins or their domains] are transformed to produce new functions, and molecular systems are combined to give rise to new, often more complex systems. … We may … deduce that [such] molecular tinkering is most probably the paradigm of molecular evolution.' Such a notion appears to contrast with the recent view of coding Alus presented by one of these authors [1].

The relative frequencies for the TE classes found by Nekrutenko and Li [1] are similar to genome-wide repeat proportions, i.e. to expectations under random sampling of sequences or random errors in predicting exons. In contrast, our findings are in good agreement with previous reports [4,25] and the above arguments that repeat-derived protein-coding sequences, especially those corresponding to Alus and other SINEs, should be rare. Indeed, Alus are derived from 7SL RNA, part of the signal recognition particle on ribosomes [23], and the strong selection for such 7SL-like secondary

0014-5793/02/$22.00
PII: S0014-5793(02)02992-7

structures that they apparently experience [24] would not leave much freedom for Alu RNAs to fulfil other roles: in particular, it would be difficult to harness them to simultaneously encode functionally important proteins.

It should be emphasized that we are not questioning the presence, or even a proposed abundance [1], of alternatively spliced transcripts containing Alus. We are questioning the notion that such transcripts will be translated to yield functional proteins, except possibly in one or two extremely rare cases. In spite of anecdotal reports of involvement of Alu-derived or -sequestered amino acid sequences in molecular recognition or binding (e.g. of the HPK1 Alu in activating AP1 [1], or of a group of Alu-derived peptides in binding tau [9]), it is not at all clear that such possibly fortuitous involvement reflects a functional role.

In summary, the available examples suggest that TEs could occasionally correspond to parts of rare, atypical proteins that arise by alternative splicing and subsequent translation, resulting in aberrant products with potentially pathological effects. In general, however, the presence of SINEs in a putative or predicted human coding sequence still appears to be a good indication that it will seldom, if ever, be translated into a functional protein in vivo.

References

[1] Nekrutenko, A. and Li, W.H. (2001) Trends Genet. 17, 619–621; Database, http://nekrut.uchicago.edu/TIG_report_140.txt.
[2] Li, W.H., Gu, Z., Wang, H. and Nekrutenko, A. (2001) Nature 409, 847–849.
[3] Brosius, J. (1999) Gene 238, 115–134; Database, http://www.ifi.uni-muenster.de/exapted-retrogenes/tables3.html, Table 2.
[4] Smit, A.F. (1999) Curr. Opin. Genet. Dev. 9, 657–663.
[5] Tugendreich, S., Feng, Q., Kroll, J., Sears, D.D., Boeke, J.D. and Hieter, P. (1994) Nature 370, 106.
[6] Claverie, J.M. and Makalowski, W. (1994) Nature 371, 752.
[7] Zietkiewicz, E., Makalowski, W., Mitchell, G.A. and Labuda, D. (1994) Science 265, 1110–1111.
[8] Iyer, L.M., Aravind, L., Bork, P. et al. (2001) Genome Biol. 2, 0051.1–11.
[9] Hoenika, J., Arrasate, M., Garcia de Yebenes, J. and Avila, J. (2002) Mol. Neurosci. 13, 343–349.
[10] Jackson, D.A., Pombo, A. and Iborra, F. (2000) FASEB J. 14, 242–254.
[11] Wilusz, C.J., Wang, W. and Peltz, S.W. (2001) Genes Dev. 15, 2781–2785.
[12] Maquat, L.E. and Carmichael, G.G. (2001) Cell 104, 173–176.
[13] Moore, M.J. (2002) Cell 108, 431–434.
[14] Iborra, F.J., Jackson, D.A. and Cook, P.R. (2002) Science 293, 1139–1142.
[15] Hilleren, P., McCarthy, T., Rosbash, M., Parker, R. and Jensen, T.H. (2001) Nature 413, 538–542.
[16] Frischmeyer, P.A., van Hoof, A., O'Donnell, K., Guerrerio, A.L., Parker, R. and Dietz, H.C. (2002) Science 295, 2258–2261.
[17] Pradet-Balade, B., Boulmé, F., Beug, H., Müllner, E.W. and Garcia-Sanz, J.A. (2001) Trends Biochem. Sci. 26, 225–229.
[18] Database http://www.ebi.ac.uk/proteome/HUMAN; Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A. et al. (2001) Nucleic Acids Res. 29, 44–48.
[19] Jurka, J. (2000) Trends Genet. 16, 418–420.
[20] Pearson, W.R., Wood, T., Zhang, Z. and Miller, W. (1997) Genomics 46, 24–36.
[21] Smit, A.F. and Green, P., RepeatMasker. http://repeatmasker.genome.washington.edu.
[22] Graur, D. and Li, W.H. (1999) Fundamentals of Molecular Evolution, 2nd edn., Sinauer, Sunderland, MA.
[23] Ullu, E. and Tschudi, C. (1984) Nature 312, 171–172.
[24] Boeke, J.D. (1997) Nat. Genet. 6, 6–7.
[25] International Human Genome Sequencing Consortium/IHGSC (2001) Nature 409, 860–921.

*Corresponding author. Fax: (39)-081-2455807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

[a] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, 16637 Prague, Czech Republic*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica, Villa Comunale, 80121 Naples, Italy*

PII: S0014-5793(02)02992-7