How incorrect annotations evolve – the case of short ORFs

Michal Linial

Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University, Jerusalem, 91904, Israel

The draft of the human genome sequence is still incomplete. The outstanding tasks include filling in some gaps, finalizing the assembly of short sequences, improving sequence accuracy and correctly identifying coding regions. However, a closely related problem that receives little attention is the substantial number of incorrect annotations that have penetrated some of the widely used databases. This article illustrates this problem using the example of ubiquitin genes, and draws some conclusions that apply to false annotations in other short open reading frames (ORFs). Although the focus is on the human genome, other genomes are equally prone to similar propagation of false annotations.

The sequence data for about one hundred complete genomes is currently available. The availability of this data increases the importance of automatic annotation methods on a genomic scale. To this end, state-of-the-art, sequence-based homology search engines are used. Consequently, information can be imported from one protein to another. Evidence for such transference is the number of proteins (~4300 proteins in humans) marked as 'similar to' or 'homologous to' a previously identified gene or protein. This process holds the risk of a substantial number of incorrect annotations propagating into, and littering, routinely used databases [1]. One such instance is the case of ubiquitin-related genes.

The case of ubiquitin

Ubiquitin is a 76-amino-acid protein found in all eukaryotes. It is one of the most conserved proteins known; its amino acid sequence is identical in humans, rodents, cattle, rabbits, pigs, chickens, frogs and other organisms [2]. Ubiquitin has several roles, the most notable of which is marking proteins for degradation. In addition, it is involved in regulating gene expression, in shaping the chromatin structure and in ribosome biogenesis. The basic 76-amino-acid unit of ubiquitin appears once or as tandem repeats, or is naturally fused to ribosomal proteins [3]. Combinations of the ubiquitin unit with other domains are also known [4].

Translating ubiquitin's coding region in its opposite orientation yields an open reading frame (ORF) of equal length but – obviously – with entirely different coding information. Using this artificial polypeptide sequence as a query in a standard protein BLAST search yields a list of high-scoring hits that could be interpreted erroneously as protein homologues. A closer inspection suggests that they are incorrectly annotated (Table 1). The potential for detecting ORFs in the noncoding strand is further amplified in the case of polyubiquitin. For example, human polyubiquitin mRNA (gi:340067) codes for a protein composed of nine tandem repeats of the basic (76-amino-acid) ubiquitin unit. Translating the protein's opposite strand provides an ORF of nearly 700 amino acids. Each of the hits listed in Table 1 corresponds to a protein with one or more ubiquitin units, as well as to an ubiquitin unit with a fused extension.

Tracing the origin of such faulty annotations brings us to the early days of sequencing effort. Specifically, *Drosophila* genes were annotated as 'polytene proteins in early-ecdysone puff 63F' [5] following translation of their noncoding strand. Once established, it is likely that BLAST searches of human DNA sequences against all translated sequences will detect those as homologues. So it seems that those faulty proteins were propagated from *Drosophila* to human and were named accordingly as 'similar to polytene protein of *Drosophila*'.

Beware the short ORFs

The case presented for the ubiquitin gene can be expanded to other short ORFs. Indeed, short sequences often yield multiple potential ORFs. Furthermore, short sequences often lack a statistically sound basis in search programs such as BLAST. From all human entries in the current nonredundant protein database, ~9% (~4660 ORFs) are very short (10 to 75 amino acids). Among those, ~770 entries are termed 'hypothetical' and the annotation for 200 of the short ORFs is based on being 'similar to' another predetermined gene or ORF.

A brief survey among short hypothetical human protein sequences (of length 10–75 amino acids) suggests that for about two-thirds of them the correspondence to an actual protein of that length cannot be currently confirmed. Although most short ORF sequences (75%) are already marked as fragments, short ORFs are often trimmed because of shifts in reading frames. Such cases are relatively easy to trace by performing a routine BLAST search. Using BLAST search against all potential translated sequences (BLASTX), it is possible to detect the occurrence of annotations in which a protein sequence is deduced from its noncoding opposite strand. In those instances, as noted for the case of ubiquitin, the fact that the protein in question is related to a well-studied protein

Corresponding author: Michal Linial (michall@cc.huji.ac.il).

http://tibtec.trends.com 0167-7799/03/$ - see front matter doi:10.1016/S0167-7799(03)00139-2

Table 1. List of human and *Drosophila* transcripts and the deduced proteins based on a noncoding strand translation

Accession GI	Protein ID XP	Locus	Name	Chromosome	Length (nt)	Sequence[a](N′-.. C′-)
18544724	087050	150938	Similar to ribosomal S27a	2p16.1	570	MSVK..GGST
18600812	086497	149330	Similar to polytene protein	1q25.3	414	MSVK..GSTA
18564004	087908	154338	Hypothetical protein	6q22.2	414	MSVK..GSTT
158146	AAA28826	M37610	Polytene protein (fly)	3L 63F5	457	EGDG..MALP
552127	AAA28821	M37605	Polytene protein (fly)	3L 63F5	229	SRVM..RWLY
18581157	090518	1447687	Hypothetical protein	13q31.1	321	MAIV..NIVN

Information obtained from NCBI Entrez server [14].
[a]Four of the amino acids at the amino (N′-) and carboxy (C′-) termini of the predicted protein are indicated.

does not guarantee the correct annotation. For example, a human 58-amino-acid protein (CTC7_HUMAN) called 'putative metallothionein C20orf127' best matches a newly identified mouse hypothetical protein XP_15014 considering the opposite strand translation. In this case, multiple evidences support the correctness of XP_15014 as an authentic protein. In another case, an ORF clearly similar to ribosomal protein L7a (Q9NU46) perfectly matched a human ORF (gi:11493425, PRO1477) of which the protein sequence had been deduced from translating the opposite strand. In this instance, the wrong ORF has already infiltrated the rat genome (gi:27706176, similar to PRO1477). Alu-derived sequences in genomes provide additional source for incorrect annotations. In the case of Alu sequences, translation in all six possible frames match predetermined ORFs. Indeed, many short hypothetical proteins are associated with the eight classes of Alu-derived sequence. An alert for such instances has been already announced [6].

The aforementioned examples of incorrect annotation make pressing the need to validate annotations on a whole-genome scale. Being 'similar to' an annotated protein appears to be insufficient. Predicting a coding gene from raw DNA sequences is based on a variety of considerations [7–9]. In the case of the highly abundant ubiquitin transcript, a search against the human EST database reveals hundreds of positive hits. For >95% of these, the coding and noncoding strand can be unequivocally defined based on the 5′ and 3′ EST tag assignment.

How can a potential user rely on the listed ORFs in the nonredundant translated database (over one million entries in April 2003) in cases that lack supporting experimental evidence? The truth of the matter is that additional criteria supporting the presence of an ORF must be included. For short ORFs, arguments based on statistical parameters – such as codon preference, synonymous and nonsynonymous substitutions, the presence of a sequence consensus (i.e. promoter, terminators, intron boundaries) – might not be reliable [9]. Unfortunately, if only a few EST sequences are available, determining the sequence directionality by 3′ or 5′ tags might not be possible. Currently, the best strategy for suggesting an ORF as coded for a genuine protein is to combine experimental evidence with a critical expert view.

Clearing up the mess

To substantiate our observation that short ORFs suffer from incorrect annotations we analyzed the confidence for correct ORFs among short ORFs from the human genome. We have eliminated those marked as 'fragment' and only tested short ORFs of length 10–75 amino acids, presumably representing full-length polypeptides. More than 40 of those short ORFs denoting 'similar to hypothetical protein' and 'hypothetical and unknown protein' were inspected using the BLASTX search engine. The level of accuracy for each ORF has been determined manually by combining the results from the search engine with available experimental evidences (unpublished results). A summary of the results is shown in Figure 1. ORFs that are likely to match their predicted sequence comprise only one-third of these sequences. Although no experimental support is available for most of them, homologues in other mammals might be used for validating correctness. The rest of the tested short ORFs are partitioned between those that suffer from one or more frame-shift mistakes, the Alu-based sequences, a wrong coding strand assignment and suspicious ORFs. Those ORFs that are characterized by a lack of any significant homology or experimental support include mistakenly annotated sequences (unpublished results) and some *bona fide* human ORFans [10]. It is important to note that, in some cases, noncoding antisense RNA molecules might be expressed naturally. Such RNA molecules are candidates for regulating gene expression by complementary strand hybridization [11]. This phenomenon should be considered when analyzing cDNA or EST supporting data.

Herein, we highlight the source of mal-annotated ORF sequences in the hope that, in the near future, their number will be reduced. Improved methods for gene prediction [12,13], manual inspection of putative ORFs

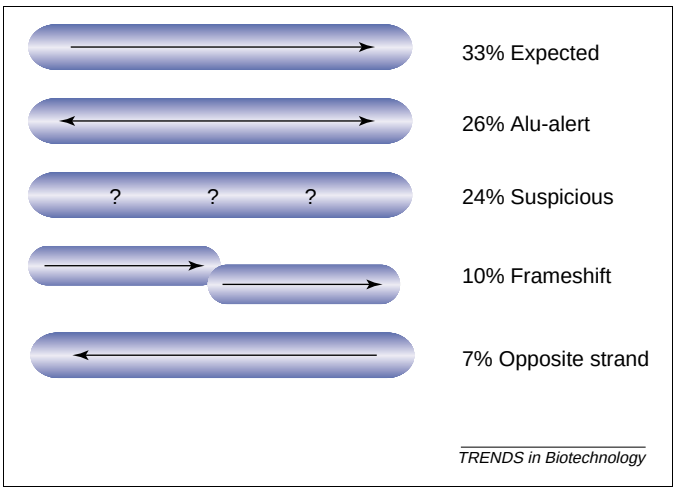

Fig. 1. Source of incorrect annotations among human short ORFs. Short ORFs (42) denoted 'hypothetical' and 'similar to other protein' were inspected. Correctness of the ORFs was marked according to the following categories: expected (forward arrow), opposite strand translation (reverse arrow), frameshift mistakes (broken arrows), Alu-alert sequences (double-headed arrow) and suspicious ORF (question marks).

and awareness that short ORFs need to be validated more carefully, combined with 'good citizenship' from the entire community, will eventually lead to a reduction in false annotated ORFs. However, faulty annotation that has already infiltrated commonly used databases tends to accumulate rather than to fade over time. An analogy between the current databases and the evolutionary process of the human genome can be drawn. Ancient contaminations in the form of pseudogenes, transposable elements and retroviruses accumulated and still decorate our genome. A systematic mechanism for cleaning up and marking potential annotation mistakes for short hypothetical ORFs is desirable. In this respect, even extensively studied proteins are vulnerable to incorrect annotation.

Acknowledgements

M.L. is a member of the HUJI-Computational Biology Center; This study is supported by the Israeli Ministry of Defense.

References

1 Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.* 17, 429–431
2 Wiborg, O. *et al.* (1985) The human ubiquitin multigene family: some genes contain multiple directly repeated ubiquitin coding sequences. *EMBO J.* 4, 755–759
3 Chan, Y.L. *et al.* (1995) The carboxyl extensions of two rat ubiquitin fusion proteins are ribosomal proteins S27a and L40. *Biochem. Biophys. Res. Commun.* 215, 682–690
4 Nenoi, M. *et al.* (2000) Interspecific comparison in the frequency of concerted evolution at the polyubiquitin gene locus. *J. Mol. Evol.* 51, 161–165
5 Izquierdo, M. *et al.* (1984) Characterization of a *Drosophila* repeat mapping at the early-ecdysone puff 63F and present in many eukaryotic genomes. *Biochim. Biophys. Acta* 783, 114–121
6 Claverie, J.M. (1992) Identifying coding exons by similarity search: Alu-derived and other potentially misleading protein sequences. *Genomics* 12, 838–841
7 Bailey, L.C. *et al.* (1998) Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* 8, 362–376
8 Reese, M.G. *et al.* (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10, 483–501
9 Carpena, P. *et al.* (2002) A simple and species-independent coding measure. *Gene* 300, 97–104
10 Fischer, D. and Eisenberg, D. (1999) Finding families for genomic ORFans. *Bioinformatics* 15, 759–762
11 Kumar, M. and Carmichael, G.G. (1998) Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.* 62, 1415–1434
12 Rigoutsos, I. *et al.* (2002) Dictionary-driven protein annotation. *Nucleic Acids Res.* 30, 3901–3916
13 Andrade, M.A. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics* 15, 391–412
14 Wheeler, D.L. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31, 28–33