Transcripts Under Selection Compose Nearly Half of the Human Genome.

Marie Sémon and Laurent Duret

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558.
Université Claude Bernard, Lyon1, 69622 Villeurbanne Cedex- France
semon@biomserv.univ-lyon1.fr

Keywords Spurious transcript, Transposable Elements, Transcriptome

Abstract

The complete sequencing of the human genome demonstrated that protein-coding regions constitute only a tiny fraction of our genome (1.5%). The amount of functional transcription units is however much more difficult to estimate, because many transcripts are spurious. We try here to evaluate the amount of sequences in the human genome that are under selective pressure to be transcribed. We built a prediction tool based on the generalized linear model, using transposable elements densities and other sequence compositional variables. We show that these features are informative enough to predict whether a sequence is transcribed or not, and - if transcribed - in which orientation. We estimate that functional transcripts constitute at least 50% of the genome, and that about one third of these transcripts do not encode proteins. The comparison of LINEs and LTR elements distribution in transcribed vs. untranscribed sequences, and on the sense vs. the antisense strand of transcripts suggests that three main factors govern selection against insertions of transposable elements in transcription units: i) selection against insertion of polyadenylation signal on the sense strand of transcription, ii) selection against insertion of promoter elements on both strands; iii) selection against the increase of the length of transcription units.

Introduction

Protein-coding regions make up 1.5% of the human genome [1]. The amount of functional transcribed sequences is however much more difficult to estimate. Thanks to transcriptome projects (e.g. ESTs), a very large number of transcription units have been identified [2]. However, some of these transcripts are functionless. Such spurious transcripts may result from the activity of cryptic promoters, e.g. originating from transposable elements or from recent pseudogenes. Some spurious transcripts may also result from the illegitimate extension of transcription, downstream of genes with weak polyadenylation signals. Contrarily to functional transcription units (FTUs), these spurious transcripts are not necessary for the proper functioning of genomes, and hence are not subject to selective pressure. The aim of this paper is to build a model to predict FTUs, and thus to evaluate the fraction of the human genome that belong to FTUs, i.e. that is under selective pressure to be transcribed.

Data set

To analyze the features that discriminate FTUs from other genomic sequences, we first had to prepare a set of sequences corresponding to known FTUs and a set of sequences that do not correspond to any known FTU. Note that our aim was to identify all FTUs, not only from protein coding genes, but also from non-coding RNA genes (ncRNA). There are presently too few known ncRNA genes to be used for such an analysis. It is however important to note that protein-genes FTUs are in fact essentially composed of non-coding sequences: in average, introns constitute 95% of the length of protein-genes FTUs [1]. We therefore decided to consider intron sequences of known protein genes as a model of FTUs, both for protein genes and ncRNA genes. As a model of non-FTU sequence, we used intergenic sequences, i.e. sequences located between transcription units annotated in databases. It should be stressed that database annotations may be incomplete, and hence that these "intergenic" sequences may in fact contain some transcription units that have not been yet identified.

We extracted from the Hovergen database [3] 2506 intronic sequences and 3753 intergenic sequences longer than 5 kb. One third of these data was isolated to be use as a test set. The rest was used for the study of compositional features and for the training of the predictive model.

Selection against insertions of transposable elements in FTUs

The most striking difference between FTUs and intergenic sequences is the distribution of transposable elements (TEs). As noted previously [4; 5], LINE L1 and LTR elements are rare in introns compared to intergenic regions. Notably, LTR elements are two times more frequent in intergenic regions than in FTUs. Moreover, LINEs and especially LTR element insertions are counter-selected on the sense strand in FTUs. LINEs and LTR elements are respectively 2 and 4 times more frequent on the sense strand of transcripts than on the antisense strand. This difference between the two strands is probably due to the fact that these TEs contain polyadenylation signals, and that insertion of such signals in the sense strand may cause the premature termination of the transcript, and hence be counterselected [4].
We further show that the proportion of LINE elements truncated at their 5' end is two times higher in FTUs than in intergenic regions, possibly because of selection against the insertion of promoter elements (located in 5' part of LINEs) within FTUs. Finally, we observed that LINE and LTR elements are shorter in FTUs than in intergenic sequences. This latter observation might reflect selection against the increase cost of transcription induced by insertion of sequences within transcription units of highly expressed genes [6].

Compositional asymmetry within FTUs

We compared the base composition (GC-content, dinucleotide content) in FTUs and intergenic regions. The only noteworthy observation is that in FTUs, the frequencies of A and G in the sense strand are respectively higher than the frequency of T and C. These AT and GC skews between the two strands are not observed in intergenic sequences. Such skews probably result from the asymmetry of the transcription process that might affect the pattern of mutation or the efficiency of DNA repair on the sense strand compared to the antisense [7; 8]. The AT and GC skews observed in FTUs, although significant, are not strong enough to be used to predict the location of transcribed regions. However, they are very reliable predictors of the orientation of transcribed regions (see below).

A model to predict FTUs

Since TE insertions are expected to be counter-selected in FTUs, but not in spurious transcripts or untranscribed region, we decided to use these features to try to differentiate FTUs from other sequences. For this purpose, we developed a prediction model (generalized linear model [9]) based on the analysis of TE distribution. We also introduced AT and GC skews in our model so that to predict the orientation of transcripts. Two predictive models, one for each orientation of transcription were trained on the learning set of FTUs and intergenic sequences. A sliding window (20 kb) is moved along the sequence, and for each window, two scores are computed, one for each transcription orientation.
The efficiency of the model to detect FTUs was first evaluated on the test set: the sensitivity and specificity of the method are both about 65%. Note that the specificity is probably underestimated because some "false positive" predictions in intergenic regions may in fact correspond to true, but unannotated, FTUs. The model performs very well to determine the orientation of transcription: 90% of the predictions correspond to the annotations, plus 5% of the sequences that are predicted to be transcribed on both strands.
We also evaluated the efficiency of the method on the 20 human ncRNA genes from the Noncoding RNAs Database that could be located on the human genome [10]. The sensitivity of the method appears to be similar for protein genes and ncRNA genes: 60% of ncRNA genes were predicted as transcribed, and 92% of them in the correct orientation.

Whole genome analysis

We then used our model on the whole human genome [1]. We used EnsEMBL annotations to locate previously known transcribed sequences. 37% of the windows contain at least one annotated transcribed region that covers more than 2 kb over 20 kb. About one half (53%) of these sequences were recognized as transcribed by the model. Moreover, 24% of the other windows are also predicted as transcribed. Given the sensitivity and specificity of the method (65%), this approach is not accurate enough to be used for automated annotation of the genome.
However, this model permits us to evaluate the part of the human genome that is transcribed under selection. Taking into account the sensitivity and specificity measured previously, our results suggest that overall, 45% of the windows contain FTUs, i.e. that about half of the human genome is under selective pressure to be transcribed. This is between the two previous estimation, namely 30% of the genome, the amount usually accepted [1], and the almost totality of the genome [11]. We found that 5% of the transcribed sequences are expressed on both strands. This observation is consistent with the literature [12], and could be biologically relevant: many non-coding RNAs are developmental regulators on the antisense strand of a coding gene (Eddy, 2001).
About one third of the predicted FTUs do not contain to any known protein-gene, and thus might correspond to ncRNA genes. As the human gene catalogue is not yet complete, and the non-coding RNAs discovery is still at it's beginning [2], the percentage we find is not outstanding. Note however that because of the size of the window (20 kb), our method can only detect relatively long FTUs. Moreover, as mentioned previously, the specificity of our method is probably underestimated. Taken together, this suggests that FTUs constitute more than 50% of our genome.

References

[1] Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., Fitzhugh W., *et al*. Initial sequencing and analysis of the human genome. *Nature* 409:860-921, 2001.

[2] Okazaki Y., Furuno M., KasukaWA T., Adachi J., Bono H., Kondo S., Nikaido I., Osato N., Saito R., Suzuki H., *et al*. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420:563-73, 2002.

[3] Duret L., Mouchiroud D. and Gouy M. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res*. 22:2360-235, 1994.

[4] Smit A.F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet. Dev*. 9:657-63, 1999.

[5] Medstrand P., Van de Lagemaat L.N. and Mager D.L. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome. Res.* 12:1483-195, 2002.

[6] Castillo-Davis CI., Mekhedov SL., Hartl DK., Koonin EV., Kondrashov FA. Selection for short introns in highly expressed genes. *Nat genet.* Aug;31(4):415-8, 2002.

[7] Duret L. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*. Dec;12(6):640-9, 2002.

[8] Green P, Ewing B, Miller W, Thomas PJ, Green ED. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet.* Apr;33(4):514-7, 2003.

[9] McCullagh P. and Nelder J.A. Generalized Linear Models.Chapman & Hall, London, 1989.

[10] Szymanski M., Erdmann V.A. and Barciszewski J. Noncoding regulatory RNAs database. *Nucleic. Acids. Res.* 31:429-31, 2003.

[11] Wong G.K., Passey D.A. and Yu J. Most of the human genome is transcribed. *Genome. Res*. 11:1975-197, 2001.

[12] Shendure J. and Church G.M. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome. Biol.* 3(9):RESEAR 0044.1-0044.14, 2002.