Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 365 (2006) 35 – 40

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

A new strategy to identify novel genes and gene isoforms: Analysis of human chromosomes 15, 21 and 22

Matteo Rè [a,1], Flavio Mignone [a,1], Michele Iacono [a], Giorgio Grillo [b], Sabino Liuni [b], Graziano Pesole [a,b,*]

[a] Dipartimento di Scienze Biomolecolari e Biotecnologie, via Celoria 26, 20133 Milano, Italy
[b] Istituto Tecnologie Biomediche, C.N.R., Sede di Bari, via Amendola 122/D, 70126 Bari, Italy

Received 10 May 2005; received in revised form 18 July 2005; accepted 27 September 2005
Available online 15 December 2005

Abstract

We present here a novel methodology for the identification of genome regions potentially spanning one or more protein coding genes. It is based on the detection of clusters of conserved sequence tags whose evolutionary dynamics, based on the observation of an excess bias of synonymous substitutions at nucleotide level and of conservative replacements at protein level, suggests a likely protein coding role. A benchmark test carried out on a 236 Mbp of human–mouse syntenic regions from human chromosomes 15, 21 and 22 identified 25 CST clusters potentially containing unannotated genes. A further annotation update of the human genome assembly revealed that 11/25 clusters actually contained a total of 20 validated genes and 10 of the remaining 14 clusters had several experimental evidence in support of the presence of protein coding genes. These findings demonstrate the effectiveness and high prediction reliability of the proposed methodology which could specifically be applied to the annotation of novel genome sequences.
Published by Elsevier B.V.

Keywords: Gene finding; Coding potential score; Conserved sequence tag; Alternative splicing; Software; Bioinformatics

1. Introduction

Although the sequence of the human genome can be considered virtually complete, we are still far from having a definitive inventory of the genes it contains. The most recent studies estimate that the actual number of human genes should range between 20,000 and 25,000 (Human Genome Sequencing Consortium, 2004), strikingly lower than the early estimates of more than 30,000 (Lander et al., 2001) or even 100,000 genes found in many textbooks.

On the other hand, the level of protein complexity has been discovered to be much higher than could be expected solely from this estimated gene number. This implies that many genes and, particularly, variants generated by alternative splicing are unknown and their expression profile has not been characterized. There is still a need for accurate tools for the analysis of genomic sequences and, in particular, for the identification of genes.

Recent comparative analyses of the human and mouse genomes suggest that nearly 5% of the genomic sequences is under active selection which is likely associated with a functional role (Waterston et al., 2002). Indeed it is well known that coding sequences, as well as many non-coding regulatory regions are under purifying selection that keeps their sequences conserved during evolution. Most of these conserved regions correspond to protein coding exons with the remaining sequences, generally denoted as "conserved non-genic sequences" (CNG) or "conserved non-coding sequences" (CNS) likely to be involved in important regulatory activities (Dermitzakis et al., 2005). However, as we still do not know either the full gene inventory, or all possible splicing isoforms specified by the human genome, we may incorrectly annotate conserved sequences not mapping to known genes as CNGs or CNSs. For this reason, we

Abbreviations: CST, Conserved sequence tag; CPS, Coding potential score.

[*] Corresponding author. Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Via Celoria, 26, 20133 Milano, Italy. Tel.: +39 02 50314915; fax: +39 02 50314912.

E-mail address: graziano.pesole@unimi.it (G. Pesole).

[1] These authors contributed equally to this work.

0378-1119/$ - see front matter. Published by Elsevier B.V.
doi:10.1016/j.gene.2005.09.041

propose the use of a more neutral term as "conserved sequence tag" (CST). Several methods have been proposed to assess the coding potential of detected CSTs. Most of these are based on the detection of a bias towards synonymous substitutions (Badger and Olsen, 1999; Rivas and Eddy, 2001; Nekrutenko et al., 2003). We have previously presented the CSTminer algorithm (Mignone et al., 2003; Castrignano et al., 2004) which has been shown to outperform other similar tools such as QRNA (Rivas and Eddy, 2001) and takes into account not only the synonymous/nonsynonymous bias but also the bias towards conservative amino acid replacements observed in real coding regions. In addition, our method does not use the generally accepted definition of conserved sequence, i.e. at least 100 bp with at least 70% identity, rather, it adopts a probabilistic criterion based on the BLAST E-value. In this way, we do not miss shorter highly conserved sequences. It should be noted that more than 10% of EMBL/GenBank annotated exons are shorter than 50 bp, with more than half of these shorter than 30 bp.

We present here a new analytical strategy for the identification of novel genes or gene isoforms based on the application of a novel version of CSTminer, specifically optimized for the identification of protein coding regions. The effectiveness of this strategy has been tested on 236 Mbp syntenic regions of human and mouse chromosome 15, 21 and 22. Results obtained proved that the large majority of predictions corresponded to newly annotated genes or had several pieces of independent evidence in support of their protein coding nature.

2. Materials and methods

2.1. Identification and characterization of CSTs

The identification of CSTs and the computation of their Coding Potential Scores (CPS) was performed with the CSTminer software. The CSTminer algorithm has been described in Mignone et al. (2003) and slightly modified in Castrignano et al. (2004). Briefly, given a pair of sequences, CSTminer identifies high scoring segment pairs (HSPs) through a Blast-like sequence comparison. The coding capacity of each CST delimited by an HSP is then assessed by assigning a coding potential score (CPS) to all CSTs showing at least 5% sequence divergence. The CPS assigned to each CST corresponds to the maximum score value obtained from each of the possible reading frames in the forward and reverse orientation.

CSTs having a CPS ≥ 7.67 were labeled as coding. This threshold provides an expected false positive rate lower than 1%.

2.2. CST Clustering

The clustering procedure described in Fig. 1 can be briefly described as follows: given a coding CST starting at nucleotide i, we calculated the number "n" of CSTs starting between nucleotides i and $i+57$ kbp (57 kbp is the average length of human genes). If n resulted ≥ 4, we labeled the group of CSTs as a "precluster". Overlapping "preclusters" were then merged to form final clusters (Fig. 1).

2.3. CST supporting features

Having defined clusters, we considered all their component CSTs and compared their absolute coordinates on the human genome golden path (NCBI.35) with the various annotations reported in UCSC database Rel. 17 including proteins, transcripts, Unigene clusters and plain ESTs, Genscan and Twinscan predictions (human genome assembly 35) as well as mouse exons (Ensembl 29.33).

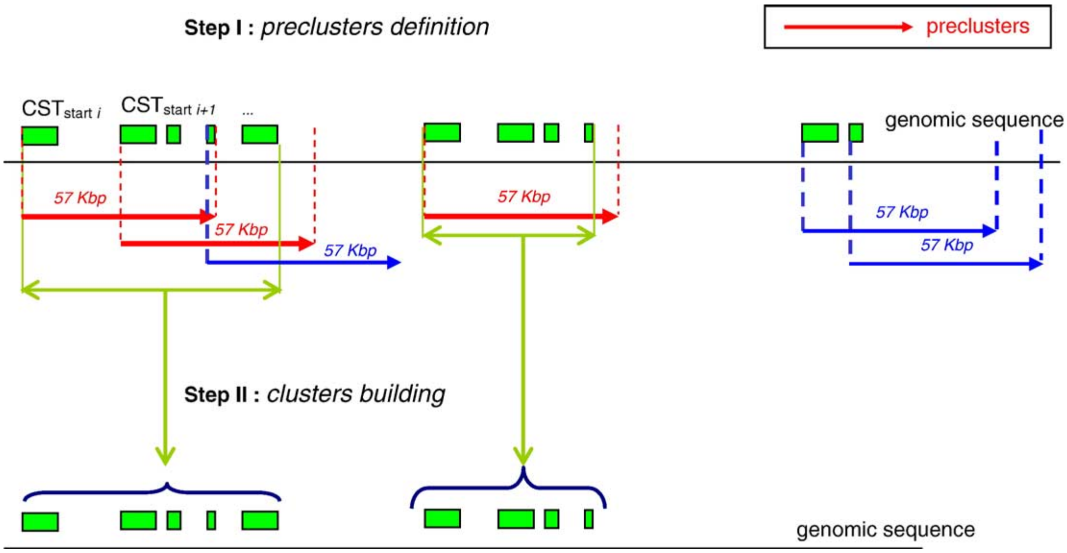


Fig. 1. Clustering procedure is described in text. (Step I) Red arrows indicate group of CSTs that satisfy pre-clustering requirements (i.e. at least 4 CSTs within 57 kbp), while blue arrows indicate CSTs which are not included into pre-clusters. (Step II) Overlapping pre-clusters are merged into a single cluster. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Results

3.1. Identification of conserved sequence tags and assessment of their coding potential

To detect conserved sequence tags (CSTs) and to assess their coding (or non-coding) nature, we applied the CSTminer program (Mignone et al., 2003) which is also available through a web interface (Castrignano et al., 2004). CSTminer compares evolutionary related sequences, identifies CSTs, and assigns to each of them a coding potential score (CPS) based on their conformation to the expected evolutionary dynamics of coding sequences at both nucleotide and amino acid levels.

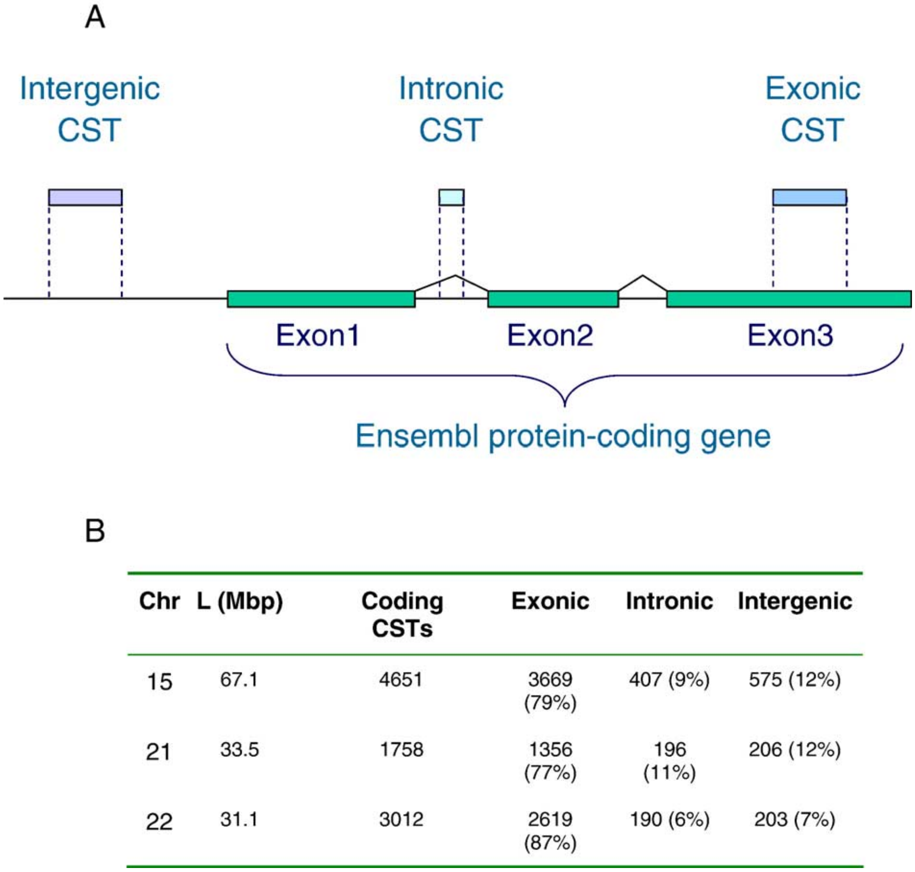
CSTs can thus be labeled as coding or non-coding depending on whether their CPS is above or below a pre-defined cutoff threshold (Castrignano et al., 2004). Nonetheless, we further refined the cutoff values here by using larger and more reliable control data sets. We used a coding data set containing 10,000 coding regions of human mouse homologous RefSeq mRNAs and a non-coding data set containing approximately 6500 18S rRNA sequences downloaded from "the European ribosomal RNA database" (http://www.psb.ugent.be/rRNA/index.html). Considering that the CPS computation relies on the analysis of the evolutionary dynamics of aligned conserved sequences (Castrignano et al., 2004), it is clear that we are unable to calculate a CPS for identical (or nearly identical) sequences. For this reason, we carried out an extensive benchmark analysis (data not shown) that defined the value of at least 5% nucleotide divergence over the whole CST length for reliable CPS computation.

3.2. Identification and characterization of CSTs in human chromosomes 15, 21 and 22

We carried out an extensive comparative analysis with CSTminer on human chromosomes 15, 21 and 22 (assembly 35) with mouse syntenic regions (assembly 33) extracting their coordinates from Ensembl Compara database (27.35).

We chose chromosomes 21 and 22 because they have been systematically analyzed (Kapranov et al., 2002; Kampa et al., 2004) and their gene inventory is well known. They thus represent a good control set to assess the sensitivity of our method. The less studied chromosome 15 is more likely to contain unannotated genes (or splicing isoforms) that could be identified by our method.

CSTminer identified more than 37,000 CSTs among which 9,421 (see Fig. 2B) were labeled as "coding" (see Materials and methods). We found a higher density of coding CSTs in



Chr	L (Mbp)	Coding CSTs	Exonic	Intronic	Intergenic
15	67.1	4651	3669 (79%)	407 (9%)	575 (12%)
21	33.5	1758	1356 (77%)	196 (11%)	206 (12%)
22	31.1	3012	2619 (87%)	190 (6%)	203 (7%)

Fig. 2. (A) Schema depicting the CST labeling procedure. (B) Classification of coding CSTs detected in chromosomes 15, 21 and 22. Each coding CST was compared with Ensembl genes and labeled as "exonic" if overlapping (minimum overlap of 20 nt) with an exon, "intronic" if overlapping with a transcript but not overlapping with any exon. Finally, we labeled a CST as "intergenic" if no overlap with any transcript was detected.

Table 1
Clusters identified following clustering procedure

Cluster_ID	Chr	Start	End	Length (bp)	Strand	CSTs (*n*)
15_m_1	15	21442791	21483252	40461	−1	6
15_m_2	15	23475925	23577468	101543	−1	23
15_m_3	15	38360938	38417350	56412	−1	24
15_m_4	15	41780732	41855832	75100	−1	19
15_m_5	15	45654614	45687571	32957	−1	4
15_m_6	15	50001320	50062488	61168	−1	18
15_m_7	15	53754992	53763425	8433	−1	6
15_m_8	15	66053263	66091139	37876	−1	4
15_m_9	15	71806395	71830629	24234	−1	5
15_m_10	15	94646202	94694584	48382	−1	4
15_p_1	15	52059066	52095395	36329	1	4
15_p_2	15	54402724	54403090	366	1	4
15_p_3	15	69190485	69281537	91052	1	5
15_p_4	15	75975490	75985491	10001	1	8
15_p_5	15	94229751	94256124	26373	1	5
15_p_6	15	99317787	99426588	108801	1	27
21_p_1	21	30886952	30937130	50178	1	8
21_p_2	21	33980361	34180645	200284	1	30
22_m_1	22	18802365	18858856	56491	−1	6
22_m_2	22	19267631	19301715	34084	−1	9
22_m_3	22	22968591	22983089	14498	−1	7
22_m_4	22	29245572	29309814	64242	−1	22
22_m_5	22	40630908	40665814	34906	−1	7
22_m_6	22	49259079	49310945	51866	−1	17
22_p_1	22	36444183	36493276	49093	1	29
Total				1315130		301

First column reports the unique identifier we assigned to each cluster, other columns show the absolute genomic position, the total length of the cluster and the total number of CSTs contained.

chromosome 15 (8.6/Mbp) with respect to chromosomes 21 and 22 (6.4/Mbp and 6.5/Mbp, respectively). The full set of coding CSTs was used for subsequent analyses. In the rest of the manuscript, “coding CSTs” will be simply called “CSTs”.

For each CST the absolute genomic coordinates were determined, allowing us to make a direct comparison with genes annotated in the Ensembl database (Rel. 28.35). Coding CSTs were then classified as “exonic”, “intronic” or “intergenic” on the basis of their location with respect to known genes as depicted in Fig. 2A.

As expected the majority of CSTs (>80%) were labeled as exonic as they corresponded with identified exons. However, we also identified 575, 206 and 203 intergenic CSTs and 407, 196 and 190 intronic CSTs in chromosomes 15, 21 and 22, respectively (Fig. 2B).

The 984 intergenic CSTs may truly correspond to as yet unannotated genes. In order to select the subset of CSTs that most likely corresponds to real genes we investigated specific features of CSTs in known coding exons. The large majority of known genes contain one or more CSTs with a modal value between four and five CSTs/gene and an average value of 8.5 CSTs/gene (data not shown). Indeed, over 80% of annotated genes contains 4 or more CSTs. In general, the CST density is much higher in gene-containing regions than in the remaining part of the genome. On average, the CST density in known gene regions is more than twofold higher than the overall average value of 7.2 CSTs/100 kbp calculated for the total syntenic regions (236 Mbp) of the three chromosomes under investigation. For this reason, we considered the observation of clustered CSTs as a typical gene signature. The density cutoff used in the present study to identify CST clusters considered genomic regions with ≥4 CSTs (the modal value of CST occurrence in real genes) spanning ≤57 kbp (the average length of annotated genes). By using a simple iterative clustering procedure (see Materials and methods), we identified 301 clustered CSTs and 25 CST clusters: 16 clusters on chromosome 15, 2 clusters on chromosome 21 and 7 clusters on chromosome 22. These clusters of coding CSTs are reliable candidates for possible novel, unannotated genes. Table 1 shows the general features of these clusters reporting for each of them absolute genome coordinates, length and number of CSTs.

In order to assess the effectiveness and reliability of the proposed methodology for gene hunting we compared the relevant intergenic regions corresponding to each of the 25 detected clusters with proteins, transcripts, Unigene representative sequences and plain ESTs mapped on the human genome as well as with Genscan (Burge and Karlin, 1997) and Twinscan (Korf et al., 2001) predictions. Indeed, the finding that

Table 2
CST clusters compared with several genomic features and gene predictions

Cluster_ID	CSTs (n)	A	B	C	D	E
		Gene predictions	Transcripts and Proteins	EST (spliced)	Mouse Ensembl exons	Human ENSG (Rel. 29)
15_m_1	6	6	1	4 (0)	6/6	1
15_m_2	23	21	22	19 (18)	17/23	0
15_m_3	24	23	22	21 (20)	14/24	2
15_m_4	19	15	17	10 (7)	18/19	2
15_m_5	**4**	**0**	**0**	**0 (0)**	**0/4**	**0**
15_m_6	18	16	12	14 (12)	13/18	1
15_m_7	6	6	6	5 (5)	6/6	0
15_m_8	4	2	0	0 (0)	2/4	0
15_m_9	5	2	1	2 (0)	0/5	0
15_m_10	**4**	**0**	**0**	**0 (0)**	**0/4**	**0**
15_p_1	4	3	3	1 (0)	1/4	0
15_p_2	**4**	**0**	**0**	**0 (0)**	**0/4**	**0**
15_p_3	5	0	1	2 (1)	0/5	0
15_p_4	8	6	3	3 (1)	7/8	0
15_p_5	**5**	**0**	**0**	**0 (0)**	**0/5**	**0**
15_p_6	27	25	21	22 (18)	25/27	2
21_p_1	8	2	3	0 (0)	4/8	3
21_p_2	30	26	28	26 (25)	28/30	0
22_m_1	6	4	3	3 (1)	3/6	2
22_m_2	9	3	2	2 (0)	9/9	0
22_m_3	7	7	7	0 (0)	7/7	0
22_m_4	22	22	17	21 (20)	22/22	2
22_m_5	7	7	6	5 (5)	3/7	2
22_m_6	17	15	15	17 (15)	15/17	2
22_p_1	29	29	29	25 (9)	9/29	1
Total	301	240	219	202 (157)	209/301	20

Columns A, B, C, D: the number indicates how many CSTs of the cluster are supported by the relevant feature. Gene predictions (A: Genscan and Twinscan predictions). Transcripts and proteins (B: RefSeq mRNAs and proteins, Trembl and Unigene sequences). EST (C: ESTs, spliced ESTs in parentheses). Mouse Ensembl exons (D: i.e. CSTs of the cluster overlapping with Ensembl gene exons in mouse genome). Human ENSG (E: number of Ensembl genes annotated in Rel. 29 with at least one exon overlapping with one CST of the cluster). Clusters not supported by any feature analyzed are in boldface.

Table 3
Intronic CSTs compared to features annotated on human genome

Intronic CSTs	
Total intronic CSTs with coding potential	800
Average length (nt)	71.7
Average CPS	8.03
EST matches	105 (13%)
Transcript matches	130 (16%)
Genscan/Twinscan predictions	86 (11%)
Total CSTs with ≥1 evidence	226 (28%)

proteins, transcripts or EST sequences mapping on the genome at coordinates overlapping with those of clustered CSTs provides significant support to the hypothesis that the relevant CSTs occur in transcribed / translated genome sequences. Analogously, CSTs overlapping with Genscan or Twinscan predictions suggest the likely presence of one (or more) genes in the genome region delimited by the CST cluster. A summary of supporting evidence for the detected CST clusters is shown in Table 2 (columns A, B and C). For 17/25 clusters, we observed supporting evidence from both transcribed sequences (ESTs and mRNAs) and gene predictions (Genscan/Twinscan), for 4/25 the supporting evidence was from either of the two criteria above, and for only 4/25 clusters were we unable to find any supporting evidence.

In a further effort to finding corroboration for coding CSTs we also checked for overlapping annotations in the mouse genome. Data shown in Table 2 (column D) show that 209/301 CSTs coincided with annotated mouse exons giving support to 19/25 clusters.

After the completion of the analysis described above a new release of Ensembl was made available (Rel. 29.35) with updated gene annotations. We compared our results to these updated annotations in order to evaluate the reliability of our approach. We observed that a total of 20 newly annotated genes mapped on 11/25 clusters (see Table 2, column E).

Of the remaining 14/25 clusters, 10/14 were supported by one or more sources of evidence (as previously described (see Table 2)) annotated in the human or mouse genome and very likely represent novel genes whereas only 4/25 remain without any kind of support and may represent false positives.

Finally we also analyzed intronic CSTs labeled as coding for their CPS as they could represent novel splice variants of known genes. Again we compared their absolute coordinates with other available annotations in human and mouse genome. Results summarized in Table 3 show that the coding nature of 226/800 intronic CSTs (28%) were supported by at least one source of additional evidence, including a highly significant subset corresponding to 108 CSTs for which all forms of supporting evidence were recovered.

4. Discussion

We present here a novel strategy capable of reliably identifying genomic regions likely to contain novel protein coding genes or gene isoforms through the application of a computational tool able to identify clusters of potentially coding conserved sequences. However, this tool has not been devised to precisely define gene or exon boundaries. In fact, in consideration of unexpectedly large fraction of overlapping genes identified in mammalian genomes (Veeramachaneni et al., 2004), it is likely that a single cluster may contain two or more gene loci. The observation of clustered CSTs, spanning the length of a typical human gene, and all showing a significant coding capacity as determined by the CPS score, should guarantee a relatively low false positive rate. Indeed, after updating of genome annotation, 11 of the 25 clusters originally predicted to overlap candidate novel genes were shown to be true positives and contained 20 newly annotated genes. We found significant supporting lines of evidence for 10 of the remaining 14 clusters, it is thus highly likely that these also overlap functional protein coding genes. This is further confirmed by the fact that 77/110 coding CSTs of these "unannotated" clusters (see Table 2) overlap known mouse exons, thus providing support for 8 more CST clusters. For example, CSTs of cluster 15_m_2 clearly match to the coding exons of ATP10A mouse gene.

For only 4 predictions (about 15%) we were unable to recover any kind of supporting evidence. These 4 were all from chromosome 15, whose annotation is less extensive than those of chromosomes 21 and 22. These clusters could correspond to false positives or to genes with a very restricted expression window and/or anomalous structure (e.g. pseudogenes). Indeed, the unusually short 15_p_2 cluster (366 bp) can be observed on the UCSC browser to map on the genome in correspondence of an annotated pseudogene (Yale pseudogene: PGO.9606.5939).

A lower amount of supporting evidence has been observed for intronic CSTs (see Table 3). In this case, we expect that higher rates of false positives as single and un-clustered CSTs are taken into account.

However, for these CSTs, experimental validation can be focused on the most reliable subset of intronic CSTs, i.e. those showing one or more supporting evidences (see Table 3), accounting for over 28% of total predicted.

As pointed out by Mathe et al. (2002) the nontrivial problem of predicting protein coding genes has to consider both alternative gene models and alternative transcripts. Non-canonical gene structures must also be taken. It is difficult to imagine a single algorithm which could take into account the possible complexity of eukaryotic genes when, for example, methods specifically designed to identify splice sites still fail on many non-canonical cases.

We believe that this goal can only be achieved using integrated approaches which split the gene finding problem into several tasks. This requires specifically designed methods for the resolution of each aspect of the problem.

A very good example of integrated approach is the Ensembl gene prediction pipeline (Curwen et al., 2004) which combines information deriving from databases and ab initio algorithms to predict and annotate genes.

From this perspective, our work demonstrates that the information provided by coding CSTs and clusters of CSTs could be integrated into Ensembl-like pipelines to improve the prediction of protein coding genes in newly sequenced genomes. In particular, CSTs could be very useful for the identification of genes with a low transcription level and for which ESTs are lacking, as well as for genes without annotated homologues. Indeed, the identification of CSTs and the clustering procedure do not rely on annotated sequences but only on the availability of orthologous/syntenous sequences for comparison.

Moreover, we have shown that clustering of CSTs has reliably identified genes that have been annotated in Ensembl only after a known sequence was made available.

We are aware that our method–like other comparative methods–could be greatly improved with a multi genomic approach, extending the analyses to three or more species. For this reason, we are currently working on the development of a new version of CSTminer suited for local multiple aligned conserved blocks.

Acknowledgements

This work was supported by STREP project "TRANSCODE" (EU), FIRB project "Bioinformatica per la genomica e la proteomica" (Ministero dell'Istruzione e Ricerca Scientifica, Italy), and by Associazione Italiana Ricerca sul cancro. We thank David Horner for critical reading of the manuscript and many helpful suggestions.

References

Badger, J.H., Olsen, G.J., 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol. Biol. Evol. 16, 512–524.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.

Castrignano, T., Canali, A., Grillo, G., Liuni, S., Mignone, F., Pesole, G., 2004. CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. Nucleic Acids Res. 32, W624–W627.

Curwen, V., et al., 2004. The Ensembl automatic gene annotation system. Genome Res. 14, 942–950.

Dermitzakis, E.T., Reymond, A., Antonarakis, S.E., 2005. Conserved non-genic sequences—an unexpected feature of mammalian genomes. Nat. Rev., Genet. 6, 151–157.

Human Genome Sequencing Consortium, E.T., 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

Kampa, D., et al., 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. 14, 331–342.

Kapranov, P., et al., 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296, 916–919.

Korf, I., Flicek, P., Duan, D., Brent, M.R., 2001. Integrating genomic homology into gene structure prediction. Bioinformatics 17 (Suppl 1), S140–S148.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Mathe, C., Sagot, M.F., Schiex, T., Rouze, P., 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30, 4103–4117.

Mignone, F., Grillo, G., Liuni, S., Pesole, G., 2003. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucleic Acids Res. 31, 4639–4645.

Nekrutenko, A., Chung, W.Y., Li, W.H., 2003. ETOPE: evolutionary test of predicted exons. Nucleic Acids Res. 31, 3564–3567.

Rivas, E., Eddy, S.R., 2001. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2, 8.

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R., Makalowska, I., 2004. Mammalian overlapping genes: the comparative perspective. Genome Res. 14, 280–286.

Waterston, R.H., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.