Prediction of human cDNA from its homologous mouse full-length cDNA and human shotgun database

Yoshifumi Fukunishi*, Harukazu Suzuki, Masayasu Yoshino, Hideaki Konno, Yoshihide Hayashizaki

Genome Science Laboratory, Tsukuba Lifescience Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Received 22 October 1999; received in revised form 1 December 1999

Edited by Takashi Gojobori

Abstract We propose a prediction method for human full-length cDNA by comparing sequence data between human genome shotgun sequence and mouse full-length cDNA. The human genome which is homologous to the mouse full-length cDNA is selected by a homology search program, and the predicted exons are connected at the exon-intron junction which gives the best homology score to the mouse full-length cDNA. The accuracy of the predicted human full-length coding region is 83.3%, and the false positive rate is 16.7%. Five human full-length proteins out of 20 proteins are correctly predicted.

Key words: Human genome; Human cDNA; Exon prediction; Mouse full-length cDNA; Shotgun sequence

1. Introduction

Many exon prediction methods have been proposed, i.e. Grail, Grail2 [1,2], Genefinder [3,4], Genscan [5,6], and these programs have succeeded especially for the exon prediction for eukaryote genome, but the prediction for full-length cDNA still remained as a difficult problem. These programs are based on a machine learning method, so the prediction accuracy is improved by the increased size of the learning data set. Currently, the exon prediction accuracy is approximately 70%, and especially the accuracy to predict an initiation codon is 40% for prokaryote genome [7]. Because the gene density of eukaryote genome is much lower than that of prokaryote genome and because we do not fully understand the mechanism of gene transcription, the prediction accuracy for eukaryote genome is expected to be low.

One of the most ambitious goals in the human genome project is to reveal ~90% of human genome until February 2000 [8–10]. High accuracy gene-finding methods are necessary to complete the annotation of the human genome. Combination of gene-finding programs and cDNA/EST data have been used for annotation. However, the prediction accuracy of the gene-finding programs is not high enough, and also sequence errors are expected for the rough draft from the high throughput sequencing in the human genome project. If the frame shift error in the open reading frame is caused by low quality sequence data, then the prediction accuracy will become lower. The number of human cDNAs is limited, because some special mRNAs may be expressed only in a restricted specific tissue: brain, fertilized egg, early embryo, etc. of which fresh tissue is not ethnically available from human.

The RIKEN mouse cDNA project has started to collect full-length mouse cDNAs expressed in early embryo, brain, etc. [11]. Currently, the number of 3′-tagged full-length mouse cDNAs is 500 000, and the sequencing of 12 000 full-length mouse cDNAs has been completed (see http://genome.rtc.riken.go.jp/index.html). The mouse gene shows a high degree of identity with the human gene. The mouse gene will be used for gene prediction without human cDNA and human EST. In this paper, we propose an exon prediction method in which the RIKEN full-length mouse cDNAs are used.

2. Materials and methods

The human cDNA sequence is predicted by our exon-intron junction prediction method for human genome DNA by use of the mouse full-length cDNA sequence. The regions on the human genome which are homologous to the mouse cDNA are suggested by a homology search program, and the regions to reproduce the human cDNA are selected afterwards. BLAST ver 1.4.10MP [12] is used for the homology search in this study.

The correspondence between the mouse cDNA sequence and the human genome DNA is neither one-to-one correspondence nor sequential. Sometimes, homologous regions are found on both plus and minus strands. Also, the same gene may often appear more than twice on one BAC sequence. We must find the true exon part of the sequences from the homologous regions including false positive. We assume that the optimal choice of human predicted exons gives the maximum homology to the template mouse cDNA sequence. All choices of predicted candidate exons are tried, and one combination of predicted exons which gives the maximum coverage is chosen as the predicted human exons. Fig. 1 is a schematic representation of mouse cDNA and homologous regions on the human genome. Regions 1–4 are homologous to the mouse cDNA suggested by homology search, and the length of each region is shown in the parentheses. Regions 1–3 are on the plus strand, while region 4 is on the minus strand of the human genome.

The program tries all options ($=16=2^4$) of these four regions, and Table 1 shows nine examples out of 16 options. The combination is represented by '0' and '1', where '0' represents the non-selected region and '1' represents the selected region. In example 1, only region 1 is selected and the length of coverage of mouse cDNA is 300 bp. In example 3, regions 1 and 2 are selected and the length of coverage is 1200 bp (= 300 bp+900 bp). This program selects the self-consistent combination, and 'NG' in the last column stands for unreasonable selection. In example 5, one region in the mouse cDNA belongs to two different regions on the human genome. In example 6, the correspondence between the mouse cDNA and the human genome is out of order. In example 7, the same troubles as in examples 5 and 6 occurred at the same time. In examples 9–15, the directions of two predicted exons are reversed. In example 16, no exon has been se-

*Corresponding author. Fax: (81)-298-36 9098.
E-mail: fukunisi@rtc.riken.go.jp

PII: S0014-5793(99)01696-8

lected. In the above examples, the combination in example 3 is the most probable combination that gives the maximum coverage of 1200 bp.

After the optimal combination is selected, the exon-intron junction is clarified according to the GT-AG rule [13]. Fig. 2 is a schematic representation of the relation between the mouse cDNA (x) and the human counterpart of exons on the human genome and the human cDNA. The upper, middle and lower lines represent the mouse cDNA, the human genome and the human counterpart cDNA, respectively. The region from a_1 to a_2 and the region from A_1 to A_2 are the homologous pairs. The region from b_1 to b_2 and the region from B_1 to B_2 are also homologous pairs. I and J on the human genome identify the true exon-intron junctions. Splice sites around the positions a_2 and b_1 are predicted by the GT-AG rule. Let i and j be one of the predicted 5′- and 3′-splice sites. A pair of partial sequences of 20 bp of the exon part around i and j are connected to form a part of the predicted cDNA sequence of 40 bp (y_{ij}). The most probable i–j pair should give the highest homology between y_{ij} and x that is the mouse cDNA sequence.

We expect that the length of CDS of the mouse cDNA is almost equal to the length of the human counterpart cDNA. To consider both conditions of the homology and the CDS length, we introduce the following score:

$$\mathrm{score}(i,j) = s(x,y_{ij}) - C\{(b_1-j) + (i-a_2) - (B_1-A_2)\}^2 \quad (1)$$

where $s(x,y_{ij})$ is a homology identity between mouse cDNA (x) and the selected partial sequence of human genome around i–j (y_{ij}), and the second term corresponds to the square of the difference of length of the mouse and predicted human CDSs, and C is a constant coefficient which is set to 1/2 for simplicity.

The most probable i–j pair is determined as the i–j pair which maximizes score(i,j). $s(x,y_{ij})$ is calculated based on a dot-matrix method.

$$s(x,y_{ij}) = \max(v(k);\ k = 1,2,3,\ldots,m-40) \quad (2)$$

where m is the length (bp) of the mouse cDNA and $v(k)$ is an overlap score between y_{ij} and the partial sequence of 40 bp between the k bp and the k+40 bp of x.

$$v(k) = \sum_{d=1}^{40} M(k+d,d) + \max\left(\sum_{d=1}^{40} M(k-n+d,d)/2 - 3|n|;\ n = -6 \sim 6\right) \quad (3)$$

$$M(i,j) = \begin{cases} 1 & \text{if } i\text{th basepair of } x \text{ is equal to } j\text{th basepair of } y_{ij} \\ 0 & \text{if } i\text{th basepair of } x \text{ is not equal to } j\text{th basepair of } y_{ij} \end{cases} \quad (4)$$

where n stands for the number of gaps in the overlap region and the gap penalty is set to −3. The first term in Eq. 3 stands for the overlap score, and the second term is added for improvement of the sensitivity if insertion/deletion exists in x or y_{ij}. M is the dot matrix of sequence x and y_{ij}.

Table 1
Combination selection method

Example	Region 4	Region 3	Region 2	Region 1	Coverage (bp)
1	0	0	0	1	300
2	0	0	1	0	900
3	0	0	1	1	1200
4	0	1	0	0	600
5	0	1	0	1	NG
6	0	1	1	0	NG
7	0	1	1	1	NG
8	1	0	0	0	900
9	1	0	0	1	NG
10	1	0	1	0	NG
11	1	0	1	1	NG
12	1	1	0	0	NG
13	1	1	0	1	NG
14	1	1	1	0	NG
15	1	1	1	1	NG
16	0	0	0	0	NG

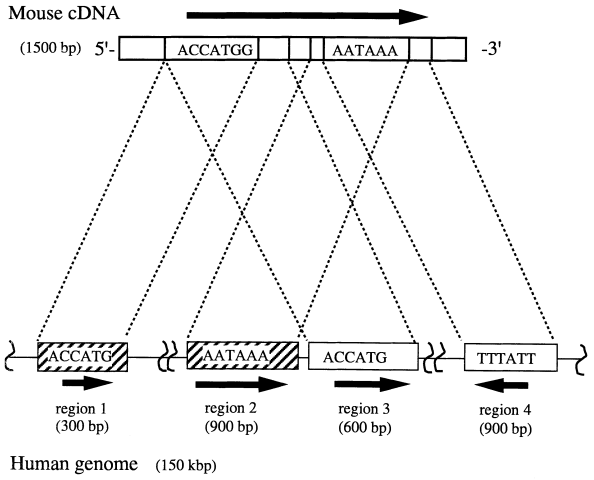

Fig. 1. Schematic representation of mouse cDNA and homologous regions on the human genome.

3. Result and discussion

Twenty mouse cDNA sequences were prepared from brain, kidney and 18-day embryo of the C57BL/6 mouse. These sequences are unfinished and will soon be published. The mouse cDNA, the predicted human cDNA and the human genome DNA used in the prediction are listed in Table 2. The global sequence identity between the human cDNA and the mouse cDNA is calculated by ALIGN [14]. Some global sequence identities are not high, because we want to examine the prediction of both homologous and low homologous cDNAs. Also, the un-translated region (UTR) of cDNA is not conserved and the lengths of mouse UTR and human UTR are different. Total human cDNA sequences of 22 340 bp and mouse cDNA sequences of 20 013 bp were used in this study.

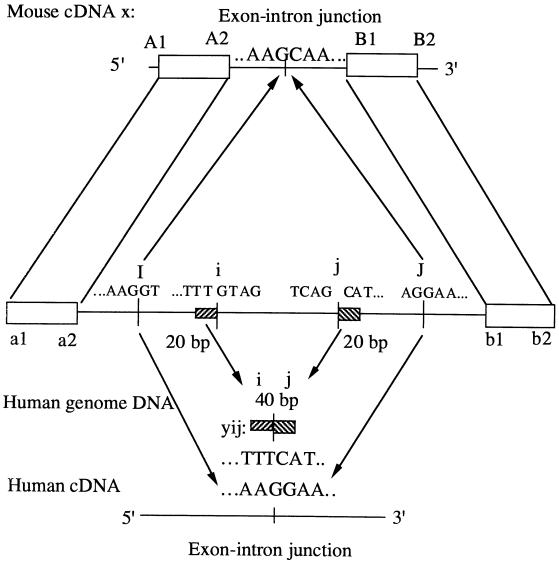

Fig. 2. Schematic representation of the relation between the mouse cDNA and the human counterpart exons on the human genome and the human cDNA.

Table 2
List of the human cDNA, the human genome DNA, and the mouse cDNA used in the prediction

Human cDNA accession no.[a]	bp[b]	Human genome accession no.[a]	Mouse cDNA	bp[c]	Global identity (%)[d]
GI4502098	1 225	AC004000	r000015l23	1 238	88.8
AF039689	1 226	HS313D11	r000019h13	1 219	83.5
HUMCIPA	1 574	AC005923	r000017m14	1 579	82.7
AF098668	1 647	HS886K2	r000016a18	1 550	81.9
HS560B094	1 157	HS560B9	r000005k05	1 110	79.8
D87292	1 137	HSE146D10	r000017i13	1 099	78.0
HSU63810	1 413	HUAC004020	r000017j20	1 327	77.2
HUMCG22	1 015	CH19F24590	r000016o07	913	74.8
HSA011497	1 207	AC003688	r000005a07	1 224	71.5
HSCALT	1 087	HSAF002995	r000017n11	1 187	71.3
HUMRAN	881	AC000097	r000005o13	1 012	67.5
GI4507370	1 883	HUMFLNG6PD	r000017k08	1 782	64.5
GI4502600	878	AB003151	r000015c07	1 168	62.7
GI4506996	1 123	AC004771	r000017m19	1 202	57.5
HSU72513	586	HSU47924	r000013h24	705	55.2
HSU65581	1 548	AC005363	r000018p19	1 036	55.0
HSU82808	1 715	HSU52111	r000016a24	1 317	53.7
HUMZC48G12	485	AC004706	r000018g20	756	52.6
AF043341	309	AC004675	r000005n01	528	50.0
AF042164	244	HS714B7	r000015d07	459	43.0
Total	22 340			22 411	

[a]GenBank accession number.
[b]Number of basepairs of human cDNA.
[c]Number of basepairs of mouse cDNA.
[d]Global homology identity between the human cDNA and the mouse cDNA.

Table 3 shows the comparison of human protein to the predicted human protein/mouse protein. The registered amino acid sequences are used for known human proteins. The local sequence identity is calculated by LALIGN [15]. In some genes (GI4506996, HSU65581, HSU82808, etc.) the size of the mouse protein is obviously shorter than the human counterpart protein. In these cases, our mouse cDNA genes do not contain the initiation codon and give the partial CDS. Thus only the C-terminus parts of these proteins show high (85.8–100%) local sequence identity.

Five human full-length proteins out of 20 proteins are correctly predicted by our method. On the other hand, Genscan ver 1.0 and Grail2 (Grailclnt ver 1.3, http://bioweb.pasteur.fr/seqanal/interfaces/grailclnt.html) are used for prediction of the

Table 3
Comparison of human protein to predicted human protein/mouse protein

Human protein accession no.[a]	aa[b]	Predicted human protein				Mouse protein	
		Global identity (%)	aa[c]	Local identity (%)	aa[d]	Global identity (%)	aa[e]
GI4502098	298	100.0	298	100.0	298	89.6	296
AF039689	303	85.8	262	99.2	262	96.7	304
HUMCIPA	480	87.6	510	94.7	473	80.2	437
AF098668	231	95.1	243	100.0	231	99.1	231
HS560B094	141	100.0	141	100.0	141	94.3	137
D87292	297	100.0	297	100.0	297	90.9	297
HSU63810	339	90.1	353	94.6	336	44.8	167
HUMCG22	193	76.6	238	100.0	187	70.1	244
HSU72513	144	73.6	108	98.1	108	38.2	128
HSA011497	211	74.9	158	100.0	158	92.4	211
HSCALT	172	67.4	116	100.0	116	95.9	172
HUMRAN	200	95.5	194	100.0	191	93.6	203
GI4507370	292	46.6	151	82.5	120	61.8	189
GI4502600	277	100.0	277	100.0	277	53.1	181
GI4506996	314	71.0	223	100.0	223	69.1	223
HSU65581	407	62.4	287	100.0	241	55.3	240
HSU82808	491	48.5	297	85.5	297	42.7	283
HUMZC48G12	123	98.4	122	98.4	122	79.7	123
AF043341	91	100.0	91	100.0	91	80.2	91
AF042164	70	84.3	70	85.5	69	56.2	80

[a]GenBank accession number.
[b]Number of amino acids of human protein.
[c]Number of amino acids of predicted human protein.
[d]Number of aligned amino acids between human protein and predicted human protein.
[e]Number of amino acids of mouse protein.

Table 4
Amino acid sequence prediction accuracy and false positive rate

	Predicted protein by our method	Predicted protein by Genscan	Predicted protein by Grail2
Prediction accuracy	83.3% (=3697 aa/4436 aa)	51.0% (=4854 aa/9517 aa)	77.9% (=3204 aa/4111 aa)
False positive	16.7% (=739 aa/4436 aa)	49.0% (=4663 aa/9517 aa)	22.1% (=907 aa/4111 aa)

same genes listed in Table 2. Genscan predicts three human full-length proteins (GI4502098, HSU63810 and AF043341) and Grail2 predicts no full-length protein.

The prediction accuracy and the false positive rate are summarized in Table 4. The prediction accuracy is the number of correctly predicted amino acids divided by the number of true amino acids, and the false positive rate is the number of false positive amino acids divided by the total number of predicted amino acids. The prediction accuracy of our program is 83.3%, and the false positive rate is 16.7%, while the prediction accuracy of Genscan and Grail2 are 51.0% and 77.9%, and the false positive rate of Genscan and Grail2 are 49.0% and 22.1%, respectively. The false positive rate of this method is half of that of Genscan. This method shows higher prediction accuracy and lower false positive rate than Grail2.

This method is based on the homology search. If the homology search did not suggest the homologous region on the human genome, such low homologous region of cDNA/protein can not be predicted. The prediction accuracy may be improved by the use of a more sensitive homology search method.

4. Conclusion

We propose a human full-length cDNA prediction method by comparing sequence data between human genome shotgun sequence and mouse full-length cDNA. The human genome which is homologous to the mouse full-length cDNA is selected by a homology search program, and the predicted exons are connected at the exon-intron junction which gives the best homology score to mouse full-length cDNA. The prediction accuracy of our method is approximately 83.3% and the false positive rate is 16.7%, while Genscan and Grail2 can predict 51.0% and 77.9% exons and the false positive rates are 49.0% and 22.1%, respectively. The false positive rate of this method is the lowest among the three programs and the prediction accuracy of this method is better than Grail2. Five human full-length proteins out of 20 proteins are correctly predicted by our method, while Genscan predicts three human full-length proteins and Grail2 predicts no full-length protein. This method is one of the most useful prediction methods developed so far.

Acknowledgements: This study has been supported by Special Coordination Funds and a Research Grant for the Genome Exploration Research Project from the Science and Technology Agency of the Japanese Government, CREST (Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation (JST), and a Grant-in-Aid for Scientific Research on Priority Areas and Human Genome Program from the Ministry of Education and Culture, Japan to Y.H.

References

[1] Uberbacher, E.C. and Mural, R.J. (1991) Proc. Natl. Acad. Sci. USA 88, 11261–11265.
[2] Mural, R.J., Einstein, J.R., Guan, X., Mann, R.C. and Uberbacher, E.C. (1991) Trends Biotechnol. 10, 66–69.
[3] Lawrence, C.B., Solovyev, V.V. and Salamov, A.A. (1994) Nucleic Acids Res. 22, 5156–5163.
[4] Lawrence, C.B. and Solovyev, V.V. (1994) Nucleic Acids Res. 22, 1272–1278.
[5] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78–94.
[6] Burge, C.B. and Karlin, S. (1998) Curr. Opin. Struct. Biol. 8, 346–354.
[7] Burset, M. and Guigo, R. (1996) Genomics 34, 353–367.
[8] Marshall, E. (1999) Science 284, 1439–1441.
[9] Marshall, E. (1999) Science 284, 1906–1909.
[10] Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L. and the members of the DOE NIH planning groups (1998) Science 282, 682–689.
[11] Sasaki, N., Nagaoka, S., Itoh, M., Izawa, M., Konno, H., Carninci, P., Yoshiki, A., Kusakabe, M., Moriuchi, T., Muramatsu, M., Okazaki, Y. and Hayashizaki, Y. (1998) Genomics 49, 167–179.
[12] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[13] Mount, S.M. (1982) Nucleic Acids Res. 10, 459–472.
[14] Myers, E. and Miller, W. (1988) CABIOS 4, 11–17.
[15] Huang, X., Hardison, R.C. and Miller, W. (1990) Comput. Appl. Biosci. 6, 373–381.