RESEARCH ARTICLE

Identification of new genes in human chromosome 3 contig 7 by graphical representation technique

Subhagata Ghosh, Amit Roy[†], Samit Adhya and A. Nandy[#,*]

Indian Institute of Chemical Biology, 4 Raja S. C. Mullick Road, Kolkata 700 032, India
[#]Present address: Environmental Science Programme, Faculty of Science, Jadavpur University, Kolkata 700 032, India
[†]Present address: Centre for Biotechnology, Shiksha Bhavana, Visvabharati University, Shantiniketan 731 235, India

The rapidly growing library of genomic length sequences and the working draft of the human genome sequence imply a concomitant need to determine new methods to analyse the sequences for rapid identification of new genes and their functions. We have developed a graphical technique for quick determination of probable coding regions in DNA sequences. In this article we apply this technique to the new sequence data from human chromosome 3 contig 7 to test the efficacy of the proposed system in a live case, and also compare with results from other genomic sequences. We report here a sampling of sequence segments that pass theoretical tests of likelihood of being genes and list several that have close homology with sequences from the expressed sequence tag (EST) databases. We also comment on the possible use of the graphical representation technique in the shotgun method of gene sequencing.

THE successful sequencing of the complete human genome[1] marks a significant step in the progress of biological sciences and an essential first step in the understanding of the functions of the human DNA. This has been possible by the rapid progress in recent years in the sequencing of various genomes brought on by the inception of the human genome project, advent of automated DNA sequencing techniques, expressed sequence tag (EST) techniques, whole genome shotgun approach and other methods. In the process, complete genomic sequences of many other organisms such as *Drosophila melanogaster*[2], yeast[3], *Haemophilus influenzae*[4], etc. have also become available. This has necessarily brought into sharp focus the necessity to find good and reliable methods for the analysis of genomic sequences to identify new genes that have not been described or identified before.

The increasingly rapid growth of the genomic sequences data bank makes the problem of gene identification even more important and researchers have to take the help of analytical computational tools that are not only fast and efficient, but are also capable of addressing new queries and parameters. The computerized search process for identification of probable gene-coding regions in a large DNA segment such as a genome can be divided into two phases as pointed out by Guigo[5]: A set of methods by which a rapid search can be made to narrow down the possibilities and focus on likely gene-coding regions, which we will refer to as Group I methods, and another, which we will call Group II methods, where more accurate search procedures can be employed to pinpoint within these selected regions the exact start and stop codons, splice sites and poly-A and other signals. Among the first group of methods that include many different ways of automatically searching and analysing DNA sequences like the Staden and the GCG methods, recent introduction of graphical representation techniques offers a radically different and quite promising way of dealing with the problem[6]. Graphical methods provide visual clues to base distribution characteristics by virtue of their representation in sequence plots[7,8] and close inspection of such plots can facilitate discrimination between gene coding and non-coding regions, especially for intron-rich sequences[9]. This is to be done with caution, however, as it has been pointed out that there are instances where the base distribution patterns provide contradictory signals, e.g. in the case of the zeta globin gene[9] and the rat myosin heavy chain gene[8], while in cases of intronless genomes such as the phage genomes, differentiating between different genes becomes a difficult task. For these reasons, it has been found that the graphical methods for exon discrimination are best done in highly intron-rich sequences where the exon regions often give strong visual clues[9]. This restriction precludes application of the graphical methods to such genomes as bacteria and the lower eukaryotes and is at this time, best applied to the human genome.

It is to be noted, however, that while the usefulness of base distribution patterns in identification of possible gene-coding regions is not in doubt, such patterns alone are not sufficient indicators of protein-coding regions in a DNA sequence[5] and more sophisticated analyses are required for final identification. Thus one has to use both Group I and Group II methods to get the best fix on probable genes. Among the Group I methods, it would appear that graphical representations of DNA sequences provide a quick and relatively easy way of determining likely coding regions, which then are to be followed up by more robust Group II formalisms. This article is part of a continuing process to determine the domains of applicability of the graphical techniques.

*For correspondence. (e-mail: anandy43@yahoo.com)

We have used the Nandy plot technique of 2D graphical representation[7] to identify and analyse a section of the human chromosome 3 contig 7 sequence and subjected the identified subsequences to some widely used Group II computational methods to identify possible protein-coding regions. While the graphical method indicates a very large number of potential candidates for possible coding regions, the Group II tools we have employed produce rather intriguing and divergent results, one confirming several potential genes with parallels in the EST libraries, while the other shows no genes in those regions at all. For comparison of the utility of the graphical technique in other species for Group I type of analysis, a summary of results of analysis of genomic segments of three species is also given.

System and methods

We have found that a graphical representation technique is useful in preliminary analysis of large DNA sequences. The object is to scan the sequence rapidly, select areas of interest and narrow down and refine the search progressively until the desired features are found. Detailed procedures and results have been reported elsewhere[7–9]. Here, we describe the main features of this technique.

Graphical representation

Briefly, a DNA sequence is represented as a series of points in a two-dimensional Cartesian plot using the following algorithm: we move one step to the left in case the base is adenine, one step up for cytosine, one step to the right for guanine, and one step down for thymine. This basically provides a running plot with the instantaneous difference between the guanine and adenine residues along the *x*-axis and that between cytosine and thymine along the *y*-axis; alternatively, one can also choose other alignments of the bases with the four cardinal directions. We name the axis system by the base name abbreviations clockwise reading from negative *x*-axis; thus the axis system described first would be called the ACGT-axis system and the two other orthogonal systems would be AGCT- and ACTG-axis systems. The cumulative effect of this representation scheme is a graph of the sequence that is characteristic of the local and global base distribution in the sequence. A study of these graphs reveals that in the case of introns with rich repetitive sequences, the plots turn out to have thin and almost unidirectional runs on the map, whereas an exon sequence with a greater admixture of all the four bases tends to concentrate the points in smaller regions, often forming dense clusters[8]. This observation provides a rough and rapid method for discriminating between possible coding and non-coding regions in new sequences[9].

Slope analysis

The slope of the plot of such a sequence can therefore be expected to show a more steady value over the intron regions, while over the exon regions the slope of the curve would vary widely. By choosing a window length of W bases and a starting point (x_j, y_j) at any convenient base number j on the sequence, Nandy[9] defined the instantaneous slope s_j at the point j as the ratio of the two displacements $u_x(j)$ and $u_y(j)$ and the average slope $\overline{S_j}$ as the average over the instantaneous displacements over the window length W:

$$\overline{S_j} = \frac{1}{W}\sum_{i=1}^{W}\frac{u_y(j+i)}{u_x(j+i)}.$$

This averaging has the effect of smoothing out instantaneous fluctuations over the window length, without disturbing the larger fluctuations occurring over longer base sequences. A plot of the average slope against the base number will differentiate between regions of small and large fluctuations, from which the regions of large fluctuations can then be subjected to closer analyses. A study of slopes of one of the axes systems will need to be supplanted by comparison with the results in other axes systems, but strong fluctuations in any one axis system can be taken as a clue to further analysis in terms of, say, cluster densities in the other systems.

Cluster density

To quantify the observed differences between the exon and intron segment representations in the 2D plot, the density of points in a segment is calculated as the number of points in the segment per unit area of the plot. From an analysis of 35 genes with around 400 introns and exons covering over a quarter million bases, Nandy[8] has shown using a frequency plot that intron representations predominantly form clusters of very low densities, and the frequency of occurrence falls off exponentially rapidly with cluster density; the exons however, grow in clustering density to 0.6 or above per unit area and then fall off gradually[8]. Subsequent work (e.g. ref. 9) also confirms this trend and, in fact, as we shall show later in this article, among the 28 coding regions identified by the GenScan software, 23 are found to have cluster densities above 0.6. For a first approximation, to identify candidates for further search for coding regions, we calculate the cluster densities of possible exon regions and accept only those that have a density of 0.6 or above, preferably in all the three axes systems.

Algorithm

These methods then give us a strategy to identify protein-coding regions:

1. Plot a graph of the entire sequence, taking small sections of say 10,000–20,000 bases at a time to identify those regions which show possibilities of forming dense clusters.
2. Do a slope analysis of the identified regions to further narrow the search to smaller segments where coding regions may exist through identification of the large fluctuations in the slope data.
3. Redo the graphical plots in smaller sections identified from step 2 to more closely identify possible dense cluster regions.
4. Calculate cluster densities of the identified regions in the three axes systems.
5. Scan the table of cluster densities to identify regions with cluster density of ~0.6 and above as starting points.
6. Subject identified regions in step 5, and, if necessary, contiguous and wider stretches of the sequence, to different Group II tools search methods and progressively narrow down the sequence segments to the best possible identification of coding regions.

Identification of such strong cluster density regions in graphical representation using the above algorithm provides for the general possibilities that (a) these may be coding regions for proteins, (b) they may constitute a part or whole of an intronless gene or a part of an exon, (c) they may be spurious signals emanating from complex base distribution patterns within intron or intra-gene segments, (d) they may be non-expressed genes, (e) they may represent pseudo-genes such as leghaemoglobins, etc. The first task is therefore to subject such regions to more rigorous analyses and then eliminate the other possibilities theoretically, and, where possible, experimentally. For example, one possible test to examine whether the identified regions have any possibility of being expressed is to search for homology with ESTs, a positive result of which will make the identified region a strong contender to being a part of an expressed gene.

Group 1 software

All the Group I software required for graphical representation analyses described above were developed in-house for use on PCs.

Group II tools

We have used several software to determine the exact locations of ORFs, splice sites, TATA regions and poly-A signals. We have used DNASIS[10] as a PC-based software to search over small selected sections of the sequence. We have extensively used the Internet to do detailed analyses of the selected regions and the neighbourhood of the sequence using the WebGene[11], ORF finder[12], BLAST search[13] and GenScan[14] procedures. Among these, GenScan comes closest to meeting the criteria set by Guigo's analysis[5].

Implementation

The new sequence, human chromosome 3 contig 7 (HC3CN7) consisting of 2,23,542 bases, was subjected to our algorithm for determination of possible protein-coding regions and thereby identify new genes. The original sequencing job was done at the Baylor College of Medicine, Houston, Texas and the final version (vs 7) added to the GenBank database on 1 April 1999, accession number AC006515. As a first step, the entire chromosome 3 contig 7 sequence as well as selected parts of it were submitted for BLAST Search but no matches were found, implying that whatever genes may exist in the sequence would be, if proved to code for proteins, new additions to the gene databases. (See, however, the concluding paragraphs of this article.)

Thus, it was necessary to submit this new sequence to a detailed analysis by other means, for which we chose first the graphical method to isolate regions that appear promising and then to submit those to a more detailed and precise search by Group II tools. A rapid search of the entire sequence viewed 20,000 bases at a time on our graphical representation showed several regions of dense clustering indicating possible coding regions. As examples, we have shown in Figure 1 *a*, the plot of one section from 1 to 20,000 nt; Figure 1 *b* shows a similar plot for the section from 60,000 to 80,000 nt. Several clustering regions of large and small sizes are observed.

Next, selected portions were subjected to slope analysis to further narrow down the possible regions where coding segments could be searched. The slope graph based on the ACGT-axis system for the region 70,000 to 80,000 nt is shown in Figure 2. It is evident that the regions around base number 71,000, 72,000 and 76,000 have several high peaks and show much more of the rapid change in slope behaviour than we expect from possible coding regions.

The next stage is to determine the cluster densities of these regions. We had remarked earlier that we expect cluster density of around 0.6 and higher to indicate possible exon regions. Among the three identified segments mentioned above in the 70,000 to 80,000 base numbers region, the ones that come close to the acceptable density ranges are sections 72,200 to 72,700 and 74,000 to 76,000. The second turned out to have cluster densities above the threshold range in all three axes systems, while the first was lower than the threshold selection limit in only one of the three systems. It is pertinent to remark that there is a trade-off necessarily made between searching for every single instance of strong clustering and broader regions extending over a few hundred bases where the clustering appears to be strong; such broad re-

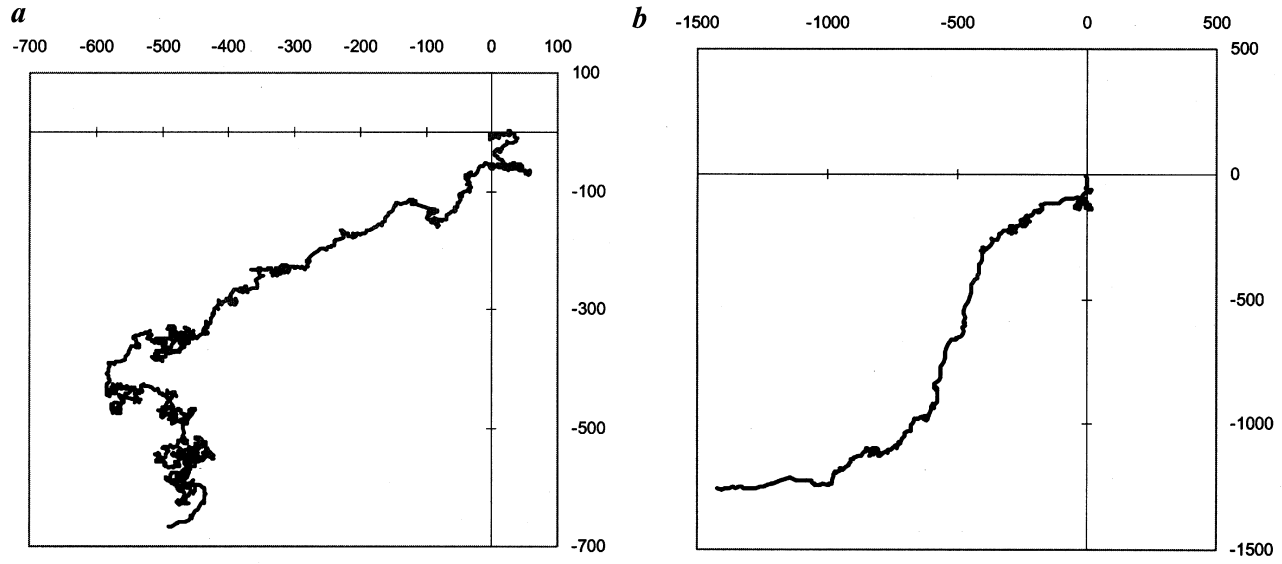

Figure 1. Plot of HC3CN7 DNA sequence from nucleotide numbers (***a***) 1 to 20,000 nt and (***b***) 60,000 to 80,000 nt. Axis system ACGT.

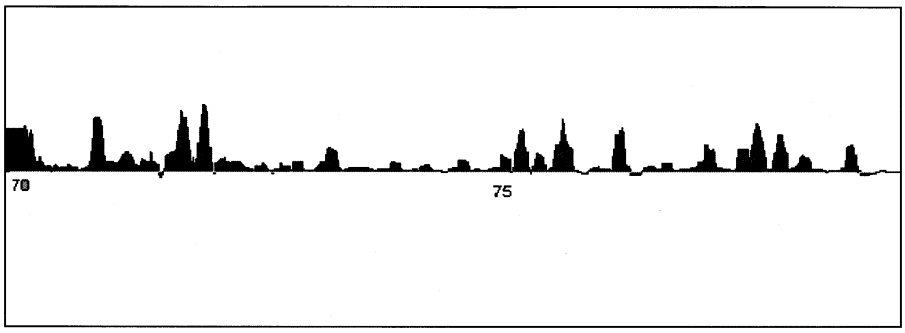

Figure 2. Slope analysis for the DNA sequence region 70,000 to 80,000 nt.

gions could arise from the presence of many small, strongly clustered points separated by short filament-like structures, but are all taken together as one group in the first approximation. The section of the human genome being investigated here shows numerous small regions of few tens of bases where the cluster density seems to be high. In this first attempt to determine coding regions we limit the attention to longer stretches of high clustering, expecting those regions to be close to or to include several exonic segments. We shall return to this point in our discussion on application of the GenScan software.

Similar analyses were done with all the other segments of the DNA sequence. Table 1 lists the regions that passed all the tests listed above. Next these are to be studied in detail to determine if they could be attributable to any gene.

We remark that since the slope analysis is done on an average basis over a window size *w*, a peak at any one point implies that the actual signal for the peak started off at least half-*w* distance ahead of the actual position. Since we have used a window size of 100 nt for our slope analysis, a peak at any one point implies the presence of signals at least 50 bases ahead. It is pertinent also to note that all exon regions do not necessarily give high cluster densities or generate running peaks in the slope analysis; there can be exon regions that do not generate such signals, or do so only lightly, while intron regions that are generally of low densities could, in some instances, generate exon-like signals[8,9].

Further tests using what we have defined earlier as Group II tools have been done on the identified high cluster-density regions to fine-tune the gene candidates.

1. The specific regions of human chromosome 3 identified from graphical representation and cluster density techniques have been analysed further by using the standard computational program DNASIS[10] to determine complete stretches between start and stop codons to identify possible intronless genes.
2. Also, we have used the WebGene genome analysis tool[11] to find out the splice sites, TATA and poly-A signals over the identified ranges, and sometimes longer stretches, to determine complete genes with exons and introns. The regions between a splice acce-

ptor site and splice donor site are subjected to standard computational programs for longest ORFs[10,12].

3. The sequence is searched for ATG, the starting amino acid codon of a protein-coding region at a distance of approximately 100 nucleotides downstream from the TATA signal.
4. Starting from ATG, ORFs between a pair of splice acceptor and splice donor sites, i.e. theoretical exons, are joined subtracting the regions between a splice donor and a splice acceptor site, i.e. theoretical introns, until a stop codon is reached.
5. After a stop codon is obtained, the poly-A signal is searched out.
6. As a final check, the GenScan software has been used to determine possible protein-coding regions and compare them with those identified by the graphical and the other Group II methods.

Table 1 lists the regions within the first 1,26,000 bases of the human chromosome 3 contig 7 DNA sequence selected by the graphical method, as candidates to be subjected to detailed analyses by Group II tools to identify possible coding segments. The choice was restricted to broad regions that pass the graphical analysis tests for acceptance in the first run. The ones marked 'Ok' passed the cluster density tests for possible candidates and were selected for detailed studies. The genes identified by the Group II tools from within the regions selected in Table 1 are given in Table 2. Comparison with the data bank on EST sequences provides a large number of candidates, as shown in Table 3. Table 4 lists the prediction made by GenScan over the entire chromosome 3 contig 7 sequence; some of the segments predicted by this software cover the regions identified in Table 1.

As a cross check on the cluster density formalism, we evaluated the densities at regions predicted as coding regions by the Group II software, but which had not been included in the preliminary analysis given in Table 1. The cluster densities of the exonic segments predicted by the WebGene software in the region 70,417–72,119 nt and in all the segments except the intronless one identified by the GenScan software are given in Table 5. In 3 out of the 4 exons of the WebGene prediction and in 23 out of the 27 exons in the GenScan prediction (leaving out one intronless gene), i.e. for over 75% of the instances, the cluster densities averaged over the three axes systems turn out to be 0.6 or above (correct to the first decimal place).

Discussion

As an example of our procedure, we consider Region 4 of Table 1 for detailed study and analysis on a PC through

Table 1. List of broad regions within the first 1,26,000 bases for the DNA sequence of human chromosome 3 contig 7 selected by graphical method for detailed analyses to identify possible coding segments. The ones marked 'Ok' passed the cluster density tests for possible candidates and were selected for detailed studies

				Cluster densities			
Start	End	Area	Bases	ACTG	AGCT	ACTG	Remark
800	1800	1368	1001	0.732	0.261	0.293	
3000	4000	3120	1001	0.321	1.009	0.308	
11,000	12,000	2756	1001	0.363	0.408	0.619	
16,050	**16,600**	**810**	**551**	**0.680**	**0.950**	**1.640**	**Ok**
21,850	**23,150**	**1200**	**1301**	**1.084**	**0.596**	**0.815**	**Ok**
25,200	25,700	1482	501	0.338	0.364	0.703	
29,500	30,700	1827	1201	0.657	0.451	0.340	
35,500	37,000	3596	1501	0.417	0.340	0.326	
39,000	40,200	957	1201	1.255	0.281	0.289	
41,000	42,700	2028	1701	0.839	0.367	0.352	
44,000	46,300	4898	2301	0.470	0.794	0.365	
47,500	48,500	1696	1001	0.590	0.820	0.439	
49,000	49,800	2419	801	0.331	0.331	0.275	
50,000	50,600	968	601	0.621	0.601	0.421	
59,300	**60,100**	**420**	**801**	**1.907**	**1.144**	**1.077**	**Ok**
62,300	63,300	2295	1001	0.436	0.472	0.798	
62,500	63,000	806	501	0.622	0.580	0.542	
72,200	72,700	660	501	0.759	0.852	0.526	
74,000	**76,000**	**1872**	**2001**	**1.068**	**0.665**	**0.616**	**Ok**
83,500	83,900	440	401	0.911	0.382	0.466	
89,800	**90,800**	**1375**	**1001**	**0.728**	**0.711**	**0.588**	**Ok**
93,100	93,500	234	401	1.714	0.532	0.472	
95,300	**95,800**	**608**	**501**	**0.824**	**0.759**	**0.589**	**Ok**
1,03,700	**1,04,000**	**368**	**301**	**0.818**	**0.896**	**0.738**	**Ok**
1,05,000	1,05,300	345	301	0.872	0.478	0.441	
1,16,500	1,26,000	24472	9501	0.388	0.174	0.304	

Table 2. List of possible genes identified within regions selected in Table 1

Regions of Table 1				Exon						
From (nucleotide no.)	To (nucleotide no.)	Possible genes	TATA signal (nucleotide no.)	From (nucleotide no.)	To (nucleotide no.)	Poly-A signal (nucleotide no.)	Gene length (nucleotides)	Protein length (amino acids)	Homology with human EST	Program used
16,000 Region 1	16,600	Gene I	14,831	14,852 15,340 15,752 16,053 16,246	14,875 15,565 15,810 16,163 16,323	16,325	510	169	–	WebGene
16,000 Region 1	16,600	Gene II	16,091	16,147	16,323	16,327	177	58	–	ORF Finder
21,800 Region 2	23,200	Gene Ia	20,932	21,033 22,060 22,364 22,873	21,083 22,107 22,392 22,954	23,130	210	69	+	WebGene
21,800 Region 2	23,200	Gene Ib	21,136	21,246 22,060 22,364 22,873	21,356 22,107 22,392 22,954	23,130	270	89	+	WebGene
21,800 Region 2	23,200	Gene Ic	21,136	21,246 22,060 22,243 22,873	21,412 22,107 22,392 22,954	23,130	447	148	–	WebGene
59,200 Region 3	60,200	Gene I	58,474	58,576 60,998 63,056 63,679 65,122 65,703	58,624 61,058 63,280 63,762 65,217 65,796	66,028	609	202	–	WebGene
73,900 Region 4	74,100	Gene I	73,844	73,908	74,162	74,300	255	84	–	DNASIS
73,900 Region 4	74,100	Gene II	68,976	69,046 69,119 69,413	69,090 69,179 69,627	69,651	321	106	+	WebGene
73,900 Region 4	74,100	Gene III	70,365	70,417 71,314 71,879 72,032	70,461 71,417 71,983 72,119	72,144	339	112	+	WebGene
89,700 Region 5	90,900	Gene I	84,334	84,369 84,870 86,582 90,249 90,740 90,890	84,431 84,932 86,632 90,300 90,787 90,948	91,165	336	111	+	WebGene
89,700 Region 5	90,900	Gene II	84,334	84,369 84,870 86,582 87,163	84,431 84,932 86,632 87,234	87,241	249	82	+	WebGene
95,200 Region 6 (complimentary strand)	95,900	Gene I	95,865	95,697	95,545	95,262	153	50	–	ORF Finder
1,03,600 Region 7	1,04,100	Gene I	1,03,177	1,0,3236 1,08,564 1,09,353 1,09,923	1,03,320 1,08,671 1,09,453 1,10,003	1,10,046 OR 1,10,063	375	125	+	WebGene

Table 3. List of human ESTs homologous with sections of possible genes in Table 2

Regions of Table 1		Homologous regions			
From To	Genes of Table 2	From To		Homologous ESTs (GenBank Acc. No.)	Type of homology
21,800 23,200 Region 2	Gene Ia, Ib, Ic	22,873	22,954	AA180807, AW970962, AA654321, T63954, T50504, H73174, H11824, AW970877, AL138096, AI636587, AA372303, AA225044, AI922231, AW838669, AI207424, R66599, AI922224, R53371	+/–
73,900 74,100 Region 4	Gene II	69,119	69,179	AA972623	+/–
73,900 74,100 Region 4	Gene III	71,314	71,417	AA229614, AA523276, AI338209, AL043486, AI038358, AW020150, BE315483, AW302017, AA807583, AA486169, AA315361	+/–
89,700 90,900 Region 5	Gene I & II	84,369	84,431	AW887544, AW936051, AW935970, AW972208, AI376269, AL043792, AL043483, AW806911, AL048277, AW994007, AW080153, AI677923, AI933808, AI829156, AI636235, AA984031, BE072336, AA906250, AA552020, AW857608, AW533462, AA378110, AA115699, BE160482, AW979039, AI739016, AI188210, AA828763, BE501866, AA631773	+/–
89,700 90,900 Region 5	Gene I	90,249	90,300	AA551123, AL036776, AW630599, W39340, W05448, T07251, BE143103, BE146869, AV625207, AV652182, AV652130, AV651994, AV646376, AV660673, AV651560, AV646369, AV646316, AW817907	+/–
1,03,600 1,04,100 Region 7	Gene I	1,08,564	1,08,671	AA808886, AA663447, AA815330	+/–
1,03,600 1,04,100 Region 7	Gene I	1,09,353	1,09,453	AA858378, AI796860, AI214337, N91295, W67780, R60134, BE160070, BE160013, AW090437, AI401741, AA815245, AW956683, AW183705, AI052817, N54352, H95603, R02194, AI125450, AI702478, AI608864, AI473420, BE010181, R09543, R49926, AW864920, AI950387, AA521181, AI889026, BE010177, AW204210, AI680287, AI422548, AI187857, T83761	+/–

DNASIS[10] and over the Internet through WebGene[11]. The DNASIS[10] scan showed a possible gene at around the 74,000 nt, details of which are given in Table 2. Further search downstream did not yield any significant candidates. However, on extending the search upstream from 74,000, several coding regions were identified by the WebGene analysis. We theoretically constructed two genes by this method. The positions of the exons (and introns by induction) identified in this region are given in Table 2.

The Gene I, Gene II and Gene III nucleotide sequences were next subjected to BLAST Search[13]. Both normal and advanced BLAST Search[13] have been carried out in human EST database, non-redundant database, GenBank and total human genome database.

The BLAST search failed to find any homologies with any existing gene in the GenBank database. The region 71,314–71,417 of chromosome 3 that includes the exon II of Region 4 Gene III, was found to be homologous with several human ESTs (Table 3). It is of interest to compare the cluster densities of the identified exon regions (Table 5) of this gene that extends from 70,417 to 72,119. It is noticed that in three out of the four identified segments, the cluster densities turn out to be close to 0.6 or higher, thus supporting our initial contention that the cluster density technique provides a quick method to identify possible protein-coding regions.

Similar analyses were carried out for all the high cluster regions identified in Table 1 through our graphical representation technique. Table 2 lists the different regions and the ORFs found in these regions. Only those candidates which have determinable TATA- and poly-A signals have been included. In some cases, there are alternative candidates for exons: Region 2 displays three and Region 5 displays two possible combinations of exons to form a gene. In Region 7, there are two candidates for the poly-A signal. In several instances, sections of the identified genes show close homologies with human ESTs (Table 3), indicating that these may be the parent genes. It is interesting to note that there is one region where the possible gene is found in the complimentary strand.

It is instructive to compare the parts of the new sequence we had originally determined to contain coding regions on the basis of cluster density and slope analysis of the corresponding 2D graphical representation: regions

Table 4. List of protein-coding regions identified by GenScan

Gn. Ex	Type	Strand	Begin	End	Length	Type of homology	Homologous ESTs (GenBank Acc. No.)
1.01	Intr	+	1931	2064	134		
1.02	Intr	+	4216	4285	70		
1.03	Intr	+	6001	6131	131	NA	NA
1.04	Intr	+	16,987	17,040	54		
1.05	Term	+	18,043	18,165	123		
1.06	PlyA	+	18,582	18,587	6		
2.02	PlyA	–	19,757	19,752	6		BQ672138, BF72647,
2.01	Sngl	–	27,858	26,956	903	+/+	AL535222,
2	Prom	–	30,932	30,893	40		BM696912, AL535483
3.11	PlyA	–	31,976	31,971	6		
3.1	Term	–	46,491	46,372	120		
3.09	Intr	–	50,477	50,303	175		
3.08	Intr	–	62,169	62,065	105	+/+	BQ447083, AI688238, AW894192
3.07	Intr	–	65,245	65,156	90		
3.06	Intr	–	65,991	65,895	97		
3.05	Intr	–	75,276	75,144	13		
3.04	Intr	–	78,475	78,315	161	+/+	BG926587, BF956166, BG152743
3.03	Intr	–	89,675	89,538	138	+/+	BG777317, BG777261, BG776785
3.02	Intr	–	98,467	98,363	105		
3.01	Init	–	99,364	99,286	79		
3	Prom	–	1,01,191	1,01,152	40		
4	Prom	+	1,04,106	1,04,145	40		
4.01	Init	+	1,08,778	1,08,973	196		
4.02	Intr	+	1,33,631	1,33,705	75		
4.03	Intr	+	1,35,342	1,35,477	136	NA	NA
4.04	Intr	+	1,35,530	1,35,703	174		
4.05	Intr	+	1,46,995	1,47,043	49		
4.06	Intr	+	1,71,331	1,71,432	102		
4.07	Intr	+	1,82,715	1,82,753	39		
4.08	Intr	+	1,96,314	1,96,529	216		
4.09	Intr	+	2,02,087	2,02,215	129		
4.1	Intr	+	2,15,088	2,15,135	48		
4.11	Intr	+	2,19,324	2,19,415	92		
4.12	Intr	+	2,21,511	2,21,571	61		

16,050 to 16,600 and 74,000 to 76,000. We found that the first one did indeed contain a possible gene replete with start and stop codons, a TATA box and possible poly-A signals. However, the second region selected for our investigation yielded only one candidate at the beginning of the segment. Follow-up investigations moving further down the sequence have provided a case for two genes, one with close homology to an EST sequence. This shows that while the graphical technique can indicate likely coding regions, it is also important to inspect closely sequence segments near the identified regions, since the clustering we observe is only indicative and not necessarily a sufficient condition for presence of coding regions. Any investigation that is based upon base distribution characteristics alone is quite likely to also select sequence segments with accidental complexity in base arrangements that could be parts of pseudogenes, remnants of old exons that are now part of non-functional introns, complexities arising out of amplification of intronic regions by deletions and accretions, and other patterns in introns and intergenic sequences[8,9].

In a complimentary exercise, we subjected selected stretches from Table 1 to analysis by the GenScan software. The results were surprisingly different from what was available from the previous attempts in that no evidence of protein-coding regions was found by GenScan. Submission of the entire sequence to GenScan resulted in four probable candidates (Table 4), the segments of some of which overlapped or were close to our original selections (16,987–17,040, 50,477–50,303, 75,276–75,144), but almost all candidates for exons scored weak to poor in terms of probability of being protein-coding regions, except for one intronless gene. A BLAST Search for ESTs yielded homologies in four exon regions predicted by GenScan.

In our initial analysis reported here, we had originally restricted our search to select broad regions of high clustering so that none of the regions in our Table 1 are below 300 bases wide, whereas the GenScan procedure has predicted exon regions extending from 39 to a maximum of 216 bases, and one intronless gene 903 bases long. A test of our procedure extending down to such small lengths of the gene shows cluster densities that meet our

Table 5. Cluster densities of coding regions

Sequence base no.			Cluster density			
From	To	Total length	ACGT	AGCT	ACTG	Average
Cluster density of coding regions identified for Gene 3 of Region 4 through WebGene						
70,417	70,461	45	0.918	0.918	1.875	0.796
71,314	71,417	104	1.444	1.000	0.693	0.648
71,879	71,983	105	0.441	0.281	0.273	0.450
72,032	72,119	88	0.733	0.547	0.423	0.568
Cluster density of coding regions identified by GenScan						
4216	4285	198	0.354	0.412	0.795	0.520
6001	6131	131	1.819	0.284	0.328	0.810
16,987	17,040	54	1.350	0.771	0.643	0.921
18,043	18,165	123	0.809	0.732	0.769	0.770
46,372	46,491	120	1.905	0.779	0.800	1.161
50,303	50,477	175	1.215	1.215	0.781	1.071
62,065	62,169	105	1.250	0.938	1.061	1.083
65,156	65,245	90	0.682	0.833	1.500	1.005
65,895	65,991	97	0.574	0.577	0.462	0.538
75,144	75,276	133	0.616	1.642	0.594	0.950
78,315	78,475	161	1.045	0.885	0.671	0.867
89,538	89,675	138	0.613	0.541	0.758	0.638
98,363	98,467	105	0.778	1.071	0.547	0.799
99,286	99,364	79	1.881	2.194	1.254	1.776
1,04,106	1,04,145	40	0.303	0.513	0.381	0.399
1,08,778	1,08,973	196	0.131	0.121	0.387	0.213
1,33,631	1,33,705	75	1.154	0.714	0.641	0.836
1,35,342	1,35,477	136	1.511	0.400	0.378	0.763
1,35,530	1,35,703	174	1.217	1.776	1.160	1.384
1,46,995	1,47,043	49	1.531	1.000	0.907	1.146
1,71,331	1,71,432	102	1.594	0.520	0.523	0.879
1,82,715	1,82,753	39	0.542	0.975	0.600	0.706
1,96,314	1,96,529	216	0.750	0.228	0.214	0.398
2,02,087	2,02,215	129	2.016	0.709	0.632	1.119
2,15,088	2,15,135	48	0.857	0.527	0.480	0.622
2,19,324	2,19,415	92	1.150	1.460	0.643	1.085
2,21,511	2,21,571	61	0.871	1.220	0.753	0.948

Table 6. Cluster density in selected genomic segments

				Coding segments tested in the region		Non-coding/intron segments tested in the region	
Species	Genome segment	Total segment length (bp)	Region tested	No. of segments	Average cluster density	No. of segments	Average cluster density
Mouse	Chromosome 8 Acc No: NW00337	4,69,521	55,013 to 1,87,664	20	0.63 ± 0.41	18	0.14 ± 0.10
A. thaliana	Chromosome 1 Acc No: AB077822	77,636	5179 to 8869 and 55,593 to 66,721	21	0.65 ± 0.09 and 0.68 ± 0.39	18	0.03 ± 0.15 and 0.29 ± 0.13
V. cholerae	Chromosome 2 Acc No: AE003853	10,72,315	1134 to 14,724	10	0.48 ± 0.16	9	0.87 ± 0.36

standards in 24 out of 31 cases (77.4%) as shown in Table 5; and implies that a detailed consideration through the graphical method of all short stretches of high clustering regions would have identified these regions along with many others.

But the differences in the predictions of WebGene and GenScan in relation to the DNA sequence under consideration are surprising. The fact that several of the WebGene predictions match with EST sequences would seem to imply that its predictions of the coding regions may not be mere artefacts, but possibly the regions so identified as complete genes could be parts of longer genes not properly identified. It is nevertheless not satisfactory that WebGene and GenScan seem to be so much at variance with each other and probably implies that in silico analyses procedures still need refining.

To test the basis of our method of selecting DNA segments as possible coding sequences if the average cluster density were to be 0.6 or higher, we tested different genomic sequences[15] for cluster densities of identified coding and non-coding sequences. The results for selected groups are given in Table 6. It is seen that the mouse genomic sequences and *Arabidopsis thaliana* cluster densities agree quite well with our hypothesis. However, bacterial genomes like that of *Vibrio cholerae* are a shade below expectations, although the non-coding regions in the selected segment give very high densities. It is to be noted, however, that our method is predicated upon the distinct differences in base organization in exon and intron segments in intron-rich sequences and hence we do not recommend our method for application to nominally intronless sequences like bacterial genomes. Mammalian and plant genomes that are rich in introns can, however, be analysed by this method to yield reasonable preliminary results to search for possible coding sequences.

It is of interest to record that our work on the human chromosome segment AC006515.7 was originally started with the previous version (AC006515.6) of the sequence submitted to GenBank on 19 March 1999, albeit with a clear warning that this was a preliminary version subject to revisions. The first run using the graphical representation technique through the sequence of 2,86,586 bases as given in this version showed a possible coding region existing in the nucleotide number range from 1 to 1000. A BLAST Search of these 1000 bases identified the sequence segment 21 to 1000 nt as an exact match with base numbers 2000 to 2979 of EMBL X15639 Phage P1 c4 *rep*L gene and base numbers 423 to 1402 of Phage P1 PP1 LREP DNA for lytic replicon containing promoter P53. Another possible coding region identified through our technique in AC006515.6 from base nos 8300 to 9100, also matched (800/801) with bacteriophage P1 *rep*L base nos 2020 to 1220. However, in the new corrected version available from 1 April 1999 (AC006515.7), the first 14,150 bases of the previous version were deleted. One can only conclude that the previous version was contaminated by extraneous genes that passed undetected in the original rapid publication process.

Comparison of patterns in the two versions of the gene sequence, AC006515.6 and AC006515.7, had clearly revealed where the two sequences started to look alike and gave the first indication that there were considerable differences in the two submissions. The availability of such visual clues makes it possible to consider using the graphical method to order the sequences generated in the shotgun method of gene sequencing. It is now the practice to have complex computer programs match the end sequences of the segments to determine overlaps and thus the continuity. In a graphical method, overlapping sequences will show up as identical patterns and therefore can be identified as possible candidates for matching the ends together. This could speed up the matching processes significantly.

Conclusion

Thus we find using our methodology that the first half of the new human chromosome 3 contig 7 sequence possibly has several candidates for protein-coding regions, which also have close homologies with human ESTs. Comparison with GenScan predictions shows that graphical searches with a finer mesh will be able to reveal even the small segments that have the potential to be considered as parts of the coding segments. From an analysis of the earlier versions of the chromosome 3 contig 7 sequence, we have noticed that the technique can also be relied upon to reveal known genes in case of accidental contamination. We thus conclude that the graphical method described briefly here and in detail elsewhere can be looked upon as a useful technique to rapidly scan large DNA sequences to determine likely coding regions, using algorithms and procedures described in detail in this article, and therefrom, with the help of Group II tools, to narrow down the search to precise locations of the coding segments.

1. http://www.nhgri.nih.gov/NEWS/sequencing_consortium.html
2. Adams, Mark D. *et al.*, *Science*, 2000, **287**, 2185–2195.
3. Goffeau, A. *et al.*, *Science*, 1996, **274**, 563-567.
4. Fleischmann, R. D. *et al.*, *Science*, 1995, **296**, 496–512.
5. Guigo, R., *Comput. Chem*., 1997, **21**, 215–222.
6. Ray, A., Raychaudhury, C. and Nandy, A., *J. Biosci*., 1998, **23**, 55–71.
7. Nandy, A., *Curr. Sci*., 1994, **66**, 309–314.
8. Nandy, A., *Curr. Sci*., 1996, **70**, 661–668.
9. Nandy, A., *Comput. Appl. Biosci*., 1996, **12**, 55–62.
10. DNASIS software, V 6.00, Hitachi Software Engineering Company Ltd, 1984, 1990.
11. WebGene Home Page, http://www.itba.mi.cnr.it/webgene/
12. ORF Finder, http://www.ncbi.nlm.nih.gov/gorf/gorf.html
13. BLAST Search, http://www.ncbi.nlm.nih.gov/BLAST/
14. GENSCAN, http://genes.mit.edu/GENSCAN.html
15. Mouse: <http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=mouse_chr.inf> *Arabidopsis thaliana*: <http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=arabid.inf> *Vibrio cholerae*: <http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=genome&gi=161>

ACKNOWLEDGEMENTS. We thank the referee for valuable suggestions, including pointing out the analysis of Guigo and co-workers that have brought great value to this work. We are grateful to Dr S. Sen of M. D. Andersen Cancer Centre, Houston, Texas for providing a preliminary version of the sequence of human chromosome 3 contig 7 in 1999 to start the analysis. We also thank Dr Chitra Dutta for providing computer and Internet facility to S.G.

Received 6 August 2001; revised accepted 3 March 2003