CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Publications of the IAS Fellows

Recognition and analysis of protein coding genes in Severe Acute Respiratory Syndrome associated Coronavirus

Ramakant Sharma[#], Jitendra Kumar Maheshwari[#], Tulika Prakash, Debasis Dash, Samir K.Brahmachari[*]

Institute of Genomics and Integrative Biology
CSIR, Mall Road, Delhi 110 007, India

# Both authors have contributed equally

*Address for Correspondence:

Prof. Samir K. Brahmachari
Institute of Genomics and Integrative Biology
(Formerly Centre for Biochemical Technology)
Mall Road, Delhi 110 007, India

Tel: 91-11-2766-7578
Fax: 91-11-2766-7471
Email: **skb@igib.res.in**

Copyright (c) 2004 Oxford University Press

Abstract:

Motivation: The recent out break of Severe Acute Respiratory Syndrome caused by SARS coronavirus has necessitated in-depth molecular understanding of the virus to identify new drug targets. The availability of complete genome sequence of several strains of SARS virus provides the possibility of identification of protein coding genes and defining their functions. Computational approach to identify protein coding genes and their putative functions will help in designing experimental protocols.

Results: In this paper a novel analysis of SARS genome using gene prediction method GeneDecipher developed in our laboratory, has been presented. Each of the 18 newly sequenced SARS-CoV genomes has been analyzed using GeneDecipher. In addition to polyprotein 1ab*, polyprotein 1a and the four genes coding for major structural proteins spike(S), small envelope (E), membrane (M), and nucleocapsid (N), 6 to 8 additional proteins have been predicted depending upon the strain analyzed. Their lengths range between 61 and 274 amino acids. Our method also suggests that polyprotein 1ab, polyprotein 1a, spike (S), membrane (M), Nucleocapsid (N) are proteins of viral origin and others are of prokaryotic. Putative functions of all predicted protein coding genes have been suggested using conserved peptides present in their ORFs.

Availability: Detailed results of GeneDecipher analysis of all 18 strains of SARS-CoV genomes are available at http://www.igib.res.in/sarsanalysis.html
Contact: skb@igib.res.in sharmark20@yahoo.co.in jkm_iitk@yahoo.co.in

Introduction

Severe acute respiratory syndrome (SARS) has emerged as a life threatening disease. Early reports on SARS appeared from China (Ksiazek *et al.,* 2003) and subsequently, cases of SARS were reported from Taiwan, Vietnam, Canada, Singapore and other countries. The range of symptoms observed in SARS affected patients are fever, dry cough, dyspnea, headache, and hypoxemia. Typical laboratory findings include lymphopenia and mildly elevated aminotransferase levels. Death may result from progressive respiratory failure due to alveolar damage (Tsang *et al.,* 2003). On an average, the mortality rate was 4%, though it varied widely according to the geographic location (WHO report, 2003) and with the strain implicated. SARS isolates from different parts of the world have been sequenced recently. Sequence analysis of nucleic acid fragments isolated from cytopathic Vero cell cultures showed that the encoded protein sequences were similar to proteins of other coronaviruses (Drosten *et al.,* 2003). However, at the nucleic acid level, no similarity was observed with any sequence in the database indicating substantial diversity. Phylogenetic analysis showed that the isolated sequence is distinct and is placed between group2 and group3 coronaviruses in the tree (Marra *et al.,* 2003).

*GeneDecipher predicts polyprotein 1ab (265…21485) in two fragments (265…13413) and (13599…21485) because there is a stop codon at location 13413. These locations are given with respect to the NCBI refseq Genome sequence.

Current computational methods like GeneMark.hmm (Lukashin and Borodovsky, 1998), Glimmer (Salzberg *et al.*, 1998), etc. face difficulty in analyzing the SARS genome due to its small size. Methods based on Hidden Markov Models (HMM) require thousands of parameters for training. This makes these methods less suitable for analyzing smaller genomes. The problem compounds in the case of SARS-CoV genomes which are about 30kb in length. Even the method most suitable for viral gene prediction till date ZCURVE_CoV (Chen *et al*., 2003) needs 33 parameters for training.

GeneDecipher originally developed for prokaryotic gene prediction, needs only 5 parameters and can therefore analyze smaller genomes too. We have trained the Artificial Neural Network on *ecoli-k12* genome coding and non-coding regions (ORFs not reported as a gene). To predict protein coding genes using GeneDecipher on viral genomes no additional training is required. This is an obvious advantage of this method over other methods. In addition it's very difficult to find negative training set (non-coding regions) for small genomes like coronavirus. Non-coding sequences for training are made by shuffling the coding sequences (Chen *et al*., 2003). The obviation of need to train specifically for the organism thus makes GeneDecipher suitable for such small genomes.

In continuation we tried to assign function to the GeneDecipher predicted SARS-CoV genes using PLHOST, a tool for functional prediction developed at our laboratory. PLHOST assigns function based upon the presence of invariant octa/hepta peptides across proteins from different species. In this paper we present the results of our analysis on 18 SARS-CoV genomes.

Methods

SARS-CoV genome sequence:

Sequences of the 18 SARS-CoV strains available in the GenBank database (http://www.ncbi.nlm.nih.gov/Entrez/genomes/viruses) were downloaded and analyzed. These include SARS-CoV Refseq (NC_004718.3),SARS-CoV TWC (AY32118), SIN2774(AY283798), SIN2748(AY283797), SIN2679(AY283796), SIN2677(AY283794), SIN2500(AY283794), Frankfurt1(AY291315), BJ04(AY279354), BJ03(AY278490), BJ02(AY278487), GZ01(AY278848), CUHKW1(AY278554), TOR2(AY274119), TW1(AY291451), BJ01(AY278488), Urbani (AY278741), HKU-39849(AY278491). Other information related to protein coding genes was retrieved from http://www.ncbi.nlm.nih.gov/genomes/SARS/SARS.html

GeneDecipher: Protein coding gene prediction software (separate manuscript communicated)

Originally GeneDecipher was developed for prokaryotic gene prediction. To execute GeneDecipher on viral genomes we prepared a heptapeptide library derived from the proteins of 56 completely sequenced prokaryotic genomes and 1096 viral genomes.

Development of GeneDecipher is based upon the observation that difference between total number of theoretically possible peptides of a given length and that which are actually observed in nature, grows drastically as this length of peptide increases. Moreover, it is interesting to note that most of these peptides selected by nature are found only in coding regions and very rarely in theoretically translated non-coding regions. This observation has prompted us to exploit this exclusivity of natural selection of peptides that are present in protein coding sequences to differentiate between coding and non-coding regions.

Prediction of a given ORF as a coding region/gene is based upon the number of heptapeptides present and the distribution of these heptapeptide along the ORF. Our output corresponding to a given ORF is a probability value (probability of this ORF being a gene). The final cut-off probability is user dependent, but it is constant for a given genome in all six reading frames (default cut-off is 0.5).

Here it is worth noting that our method is independent of any other evidences, e.g. ribosome binding site signals (in order to prove the strength of the hypothesis) such kinds of constraints are being used by various existing methods

The method can be divided into five major steps (Figure1):

1. Generation of a peptide library.
2. Artificial translation of a given genome into 6 reading frames
3. Conversion of each translated sequence into an integer coded sequence. (one corresponding to each reading frame)
4. Training of artificial neural network (ANN).
5. Deciphering genes using trained ANN.

Detailed description of the method has been provided in GeneDecipher manuscript ID: BIOINFO-2003-0492

PLHOST: Function Assignment Tool (Manuscript communicated)

We used PLHOST (Peptide Library based Homology Search Tool) for the identification of invariant peptides which serve as functional signatures from completely sequenced genomes. (Brahmchari & Dash, 2001).

The algorithm generates organism specific libraries of octa/hepta peptides from all proteins of selected genomes. Redundant peptides are removed from each library. These

peptide libraries are then compared with each other to note all octa/hepta peptides present invariantly across a specified minimum number of genomes. Overlapping octa/hepta peptides are back stitched to generate longer conserved peptides which occur in functionally similar proteins, hence called functional signatures.

Detailed Description of the method has been given PLHOST manuscript ID: BIOINFO-2003-0496.

Results and Discussion:

A systematic sensitivity and specificity analysis of GeneDecipher has been done on 10 microbial genomes (Figure2). Further analysis of GeneDecipher on viral genomes is presented here.

Testing of GeneDecipher on viral genomes:

To test our method on viral genomes we first analyzed *Human Respiratory Syncytial Virus (HRSV),* complete genome using GeneDecipher. Comparison of GeneDecipher results with state of the art method ZCURVE_CoV has been done (Table1). ZCURVE_CoV is able to predict 8 annotated proteins out of 11 reported at NCBI without any false positives. ZCURVE_CoV was unable to predict the following three genes: PID 9629200 (location 626…1000, non-structural protein2 (NS2)); PID 9629205 (location 4690…5589, attachment glycoprotein (G)); and PID 9629208 (location 8171…8443, matrix protein 2(M2)). GeneDecipher predicted 10 out of total 11 annotated proteins of HRSV without any false positives. The gene missed by GeneDecipher was PID 9629208 (location 8171…8443, matrix protein 2) which was notably missed by ZCURVE_CoV too.

This successful prediction of protein coding regions in *HRSV* genome increases our confidence to predict protein coding regions on newly sequenced SARS-CoV genomes.

Analysis of SARS-CoV using GeneDecipher:

We analyzed all 18 strains of SARS-CoV using GeneDecipher. (Detailed results are available on the website given above). GeneDecipher predicts a total of 15 protein coding regions in SARS-CoV genomes including both the polyproteins 1a, 1ab (Sars2628 C-terminal end of Polyprotein1ab), and all four known structural proteins (M, N, S, and E) for each of the 18 strains. GeneDecipher also predicts 6 to 8 additional coding regions depending on the genome sequence of the strain used. The length of these additional coding regions varied between 61 and 274 amino acids.

GeneDecipher predicts 12 coding regions which are common to all 18 strains (Table2), and one coding region (Sars63, sars6 at NCBI refseq genome) present in 5 strains.

GeneDecipher predicts gene Sars90 in GZ01 strain, and Sars154 (Sars 3b at NCBI refseq genome) in BJ02 strain specifically.

These 12 common protein coding regions consist of the 6 basic proteins of SARS-CoV (2 polyproteins and the 4 structural proteins); Sars274 (Sars3a at NCBI refseq database), Sars122 (Sars7a at NCBI refseq database), Sars78 (already reported with start shifted as ORF14/Sars9c in TOR2 strain) ; and three newly predicted (false positives with respect to current annotation at NCBI) protein coding regions Sars174, Sars68, and Sars61. The three newly predicted genes lie completely within polyprotein 1a genomic region. Although our method discards such genes in bacterial genomes, possibility of finding such genes in viral genomes has not been ruled out. As these genes are present in all 18 strains it is likely that they are protein coding genes.

We predict three more coding regions Sars63, Sars154, and Sars90 apart from the 12 discussed above. Sars63 is identified in 5 strains and not identified in remaining 13 strains. This coding region is already reported in NCBI refseq (Sars6). Here we can not comment much about the existence of Sars63 (Sars6 at NCBI refseq) because it is identified in 5 strains and not identified in rest 13. This is due to high density of non-synonymous mutations across strains in this region. Two coding regions Sars154 (sars3b at NCBI), and Sars90 (newly predicted in GZ01 starin) are identified in only one strain. Since these two coding regions are identified in only one strain, they are less likely to be protein coding regions, as also suggested by ZCURVE_CoV (Chen *et al.,* 2003) analysis. The locations of these three genes in different strains are provided in Table 3.

Since the peptide libraries are made from the genome sequences of various organisms, the evolutionary origin of a given protein can be traced. If the protein is rich in heptapeptides found occurring in viral genomes then that protein is considered to be of viral origin. We found that 5 core proteins (two polyproteins and three structural proteins M, N, and S) are of viral origin. The remaining, including 3 new predictions, are of prokaryotic origin. It is interesting to that from the same DNA region we are getting proteins in different frames which contain peptides from different origin. Here, how same DNA sequence can code for both bacterial and viral origin is intriguing. This might explain why these new protein coding genes were not detected in primary attempts based on homology to other known viral genome sequences.

Comparison with the existing system - ZCURVE_CoV:

Comparison of GeneDecipher, ZCURVE_CoV results with the known annotations for Urbani and TOR2s trains of SARS-CoV are presented in tables 4a and 4b.

In general, GeneDecipher results are in good agreement with the known annotations. In case of Urbani strain GeneDecipher predicts all the known genes except Sars84(X5), Sars63(X3) and Sars154(X2). Sars84(X5) and Sars63(X3) are supported by ZCURVE_CoV whereas Sars154(X2) is missed by both the methods. GeneDecipher predicts four new genes in this strain which incidentally are not supported by

ZCURVE_CoV. It is noticeable that out of these four genes Sars78 is already known for strain TOR2 as ORF14/Sars9c. This supports the likelihood of the gene being present in Urbani strain. However, ZCURVE_CoV predicts 2 new genes which are not supported by GeneDecipher either.

GeneDecipher predictions for TOR2 strain are identical with those for Urbani strain. In this strain GeneDecipher predicts 9 known genes but fails to predict 6 genes with known annotations. These 6 genes are: Sars154 (ORF4), Sars98 (ORF13), Sars63 (ORF7), Sars44 (ORF9), Sars39 (ORF10), and Sars84 (ORF11). Of these, Sars154 (ORF4) and Sars98 (ORF13) are also missed by ZCURVE_CoV. It is to be noted that both Sars44 (ORF9) and Sars39 (ORF10) are ORFs very small in length (44 and 39 amino acids respectively), and their presence too is not consistent across various SARS strains. Sars63 (ORF7) has been predicted by GeneDecipher in 5 other strains but not in the two strains considered here.

Mutation Analysis:

Analysis using multiple sequence alignment (ClustalW) for 3 newly predicted protein coding genes Sars174, Sars68 and Sars61 across all 18 strains shows:

1. Sars68 has one point mutation at location 80 GAT->GGT (D->G) Sin2677 strain.
2. Sars174 has two synonymous point mutations at location 204 CGA->CGC in GZ01 strain and at location 447 CTG->CTT in BJ04 strain.
3. Sars61 has one point mutation at location 119 CTG->CAG (L->Q) in GZ01 strain.

These three newly predicted genes are present in all 18 strains without significant mutations and has no significant hits with BLASTP in non-redundant database. This indicates that these three proteins might have crucial biological functions specific to SARS-CoV. Therefore these coding sequences might serve as candidate drug targets against SARS.

Function Assignment:

In total we predict 15 coding regions in SARS-CoV out of which functions of the four structural proteins (M, N, S and E) have already been assigned. Although the polyprotein 1ab has been assigned only replicase activity, our analysis implies that the replicase activity is associated with Sars2628 (C terminal of ORF 1ab) fragment. The complete 1ab polyprotein contains 6 functional signatures of which polyprotein 1a contains signatures associated with metabolic enzymes (Table 5a). Functions were assigned to the polyproteins on the basis of peptides (length 7 or more amino acids) occurring in proteins having similar functions in at least 5 different organisms. Other predicted genes/protein coding regions contain peptides which occur in fewer genomes. Based on these peptides we suggest functions, albeit with lesser confidence (Table 5b). The biological relevance of these finding remains to be explored.

Conclusion:

In this paper we have predicted 4 new genes including Sars78 (already known in TOR2 strain) in SARS-CoV. Our analysis also corroborates the finding of ZCURVE_CoV (Chen *et al.*, 2003) that ORF Sars154 (listed in Refseq as Sars3b) is unlikely to be a coding region. We have also assigned functions to the two polyproteins 1ab and 1a. In addition to replication associated function of C-terminal of 1ab polyprotein, our analysis implies that the polyprotein 1a may be associated with metabolic enzyme like functions. In all, six peptide signatures are present in polyprotein 1ab. We have suggested putative function for other 9 proteins including ones newly predicted by GeneDecipher.

Acknowledgement: We acknowledge Dr. S. Ramachandran and Dr. Bina Pillai for invaluable discussions and Pankaj Bhatnagar for technical support. We also acknowledge Council of Scientific and Industrial Research for financial support.

References

Brahmachari, S.K. and Dash, D. (2001) a computer based method for identifying peptides useful as drug targets. PCT international patent publication (WO 01/74130 A2, 11th October 2001).

Chen,L., Ou,H., Zhang,R. and Zhang,C. (2003) Z-CURVE_CoV: a new system to recognize protein coding genes in coronavirus, and its applications in analyzing SARS-CoV genomes *Biochemical and Biophysical Research Communications,* **307**,382-8.

Cumulative number of reported cases of severe acute respiratory syndrome (SARS). Geneva: World Health Organization, 2003. (Accessed April 9, 2003 at http://www. who.int/csr/sarscountry/2003_04_04/en/.)

Drosten,C., Günther,S. and Preiser,W., (2003) Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N Engl J Med.*, (www.nejm.org on April 10, 2003.)

Ksiazek,T.G., Dean Erdman,P.H. and Goldsmith,C.S. (2003) A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N Engl J Med,* **348**, 1947-58.

Lukashin., A.V. and Borodovsky, M.(1998) GeneMark.hmm: New solution for gene finding *Nucleic Acid Research,* **26**, 1107-15.

Marra,M.A., Jones,S.J., Astell,C.R., Holt,R.A., Brooks-Wilson,A. (2003) The Genome sequence of the SARS-associated coronavirus. *Science,* **300**, 1399-404.

Mathe,C., Sagot,M.F., Schiex,T. and Rouze, P. (2002) Current Methods of gene prediction their strength and weaknesses. *Nucleic Acid Research*, **30**, 4103-17.

Salzberg,S.L., Delcher,A.L., Kaif, S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acid Research,* **26**, 544-8.

Tsang,K.W., Ho,P.L. and Ooi,G.C., (2003) A cluster of cases of severe acute respiratory syndrome in Hong Kong. *N Engl J Med*, **348**, 1977-85.

Table1. Comparison of GeneDecipher results with ZCURVE_CoV results on HRSV genome, with respect to annotated genes

Annotated genes			ZCURVE_CoV			GeneDecipher		
Start	End	Length	Start	End	Length	Start	End	Length
99	518	139	99	518	139	99	518	139
626	1000	124	--	--	--	626	1000	124
1140	2315	391	1140	2315	391	1140	2315	391
2348	3073	241	2348	3073	241	2348	3073	241
3263	4033	256	3158	4033	291	3158	4033	291
4303	4500	65	4303	4500	65	4303	4500	65
4690	5589	299	--	--	--	4690	5589	299
5666	7390	574	5666	7390	574	5621	7390	589
7618	8205	195	7618	8205	195	7618	8205	195
8171	8443	90	--	--	--	--	--	--
8509	15009	2166	8443	15009	2188	8443	15009	2188

Table2: Protein coding genes predicted by GeneDecipher in SARS-CoV Refseq common to all 18 strains.

S.No.	Start	Stop	Frame	Length		Feature
				bp	aa	
1	265	13413	1+	13149	4382	Sars1a polyprotein
2	701	1225	2+	525	174	Sars174(new prediction)
3	1397	1603	2+	207	68	Sars68(new prediction)
4	8828	9013	2+	186	61	Sars61(new prediction)
5	13599	21485	3+	7887	2628	Sars2628(C-terminal end of polyprotein 1ab)
6	21492	25259	3+	3768	1255	Spike (S) protein
7	25268	26092	2+	825	274	Sars274(Sars 3a)
8	26117	26347	2+	231	76	Sars76(Sars4)
9	26398	27063	1+	666	221	Sars221(Sars5)
10	27273	27641	3+	369	122	Sars122(Sars7a)
11	28120	29388	1+	1269	422	Sars422(Sars9a)
12	28559	28795	2+	237	78	Sars78 (Identical to ORF 14/Sars9c in TOR2 with shifted start)

Table3: Identification of Sars90, Sars63, Sars154 as protein coding genes by GeneDecipher in various strains of SARS-CoV

S.No.	Strain name	Sars90 (New prediction)	Sars63(Sars6 at NCBI)	Sars154(Sars 3b at NCBI)
1	SARS_2748	--	--	--
2	SARS_bj01	--	27055..27246	--
3	SARS_bj02	--	27074..27265	25689..26153
4	SARS_bj03	--	27070..27261	--
5	SARS_bj04	--	27058..27249	--
6	SARS_frankft1	--	--	--
7	SARS_urbani	--	--	--
8	SARS_gz01	24492..24764	27058..27249	--
9	SARSsin2500	--	--	--
10	SARS_sin2677	--	--	--
11	SARS_sin2679	--	--	--
12	SARS_sin2774	--	--	--
13	SARS_chuk	--	--	--
14	SARS_tw1	--	--	--
15	SARS_twc	--	--	--
16	SARS_hku39849	--	--	--
17	SARS_refseq	--	--	--
18	SARS_TOR2	--	--	--

Table 4(a). Comparison of GeneDecipher results with ZCURVE_CoV results on SARS-CoV genome Urbani strain, with respect to annotated genes

Annotated genes			ZCURVE_CoV			GeneDecipher			Features
Start	End	Length	Start	End	Length	Start	End	Length	
265	13398	4377	265	13398	4377	265	13413	4382	ORF 1a
--	--	--	--	--	--	701	1225	174	Sars174(New prediction by GeneDecipher)
--	--	--	--	--	--	1397	1603	68	Sars68(New prediction by GeneDecipher)
--	--	--	--	--	--	8828	9013	61	Sars61(New prediction by GeneDecipher)
13398	21485	2695	13398	21485	2695	13599	21485	2628	ORF 1b
21492	25259	1255	21492	25259	1255	21492	25259	1255	S protein
25268	26092	274	25268	26092	274	25268	26092	274	Sars274(X1)
25689	26153	154	--	--	--	--	--	--	Sars154(X2)
26117	26347	76	26117	26347	76	26117	26347	76	E protein
26398	27063	221	26398	27063	221	26389	27063	224	M protein
27074	27265	63	27074	27265	63	--	--	--	Sars63(X3)
27273	27641	122	27273	27641	122	27273	27641	122	Sars122(X4)
--	--	--	27638	27772	44	--	--	--	Sars44
--	--	--	27779	27898	39	--	--	--	Sars39
27864	28118	84	27864	28118	84	--	--	--	Sars84(X5)
28120	29388	422	28120	29388	422	28120	29388	422	N protein
--	--	--	--	--	--	28559	28795	78	Sars78(Identical to ORF 14/Sars9c in TOR2 with shifted start)

Table 4(b). Comparison of GeneDecipher results with ZCURVE_CoV results on SARS-CoV genome TOR2 strain, with respect to annotated genes

Annotated genes			ZCURVE_CoV predicted genes			GeneDecipher predicted genes			Features
Start	End	Length	Start	End	Length	Start	End	Length	
265	13398	4377	265	13398	4377	265	13413	4382	ORF 1a
--	--	--	--	--	--	701	1225	174	Sars174(New prediction by GeneDecipher)
--	--	--	--	--	--	1397	1603	68	Sars68(New prediction by GeneDecipher)
--	--	--	--	--	--	8828	9013	61	Sars61(New prediction by GeneDecipher)
13398	21485	2695	13398	21485	2695	13599	21485	2628	ORF 1b
21492	25259	1255	21492	25259	1255	21492	25259	1255	S protein
25268	26092	274	25268	26092	274	25268	26092	274	ORF3(Sars274)
25689	26153	154	--	--	--	--	--	--	ORF4(Sars154)
26117	26347	76	26117	26347	76	26117	26347	76	E protein
26398	27063	221	26398	27063	221	26389	27063	224	M protein
27074	27265	63	27074	27265	63	--	--	--	Sars63(ORF7)
27273	27641	122	27273	27641	122	27273	27641	122	Sars122(ORF8)
27638	27772	44	27638	27772	44	--	--	--	Sars44(ORF9)
27779	27898	39	27779	27898	39	--	--	--	Sars39(ORF10)
27864	28118	84	27864	28118	84	--	--	--	Sars84(ORF11)
28120	29388	422	28120	29388	422	28120	29388	422	N protein
28130	28426	98	--	--	--	--	--	--	ORF13
28583	28795	70	--	--	--	28559	28795	78	Sars78(Identical to ORF 14/Sars9c in TOR2 with shifted start)

Table 5(a): Functional assignment of polyproteins in SARS (Urbani) Genome using PLHOST

S.No.	NCBI annotation	Conserved peptide signature	Function assigned
1	Sars1ab (Poly protein 1ab)	RIRASLPT	Phosphoglycerate kinase
		RSETLLPL	Sulfite reductase (NADPH), Flavoprotein beta subunit
		LDKLKSLL	Probable acyl-CoA thiolase
		ATVVIGTS	cell division protein ftsZ
		NVAITRAK	DNA-binding protein, probably DNA helicase
		LQGPPGTGK	DNA helicase related protein
2	Sars 1a poly protein 1a	RIRASLPT	Phosphoglycerate kinase
		RSETLLPL	Sulfite reductase (NADPH), Flavoprotein beta subunit
		LDKLKSLL	Probable acyl-CoA thiolase
3	Sars 2628 (C terminal of Sars 1ab)	ATVVIGTS	cell division protein ftsZ
		NVAITRAK	DNA-binding protein, probably DNA helicase
		LQGPPGTGK	DNA helicase related protein

Table5 (b): Suggested functions for some of the non-structural genes in SARS-CoV using PLHOST

S.No.	Gene	Peptide Signature	Suggested function
1	Sars174(new prediction)	TLSKGNAQ	ABC transporter ATP binding protein *[Lactococcus lactis subsp. lactis]*
		VAQMGTLL	Cytochrome c oxidase folding protein *[Synechocystis sp. PCC 6803]*
2	Sars68(new prediction)	LVLVLILA	*putative major facilitator superfamily protein [Schizosaccharomyces pombe]*
		TQTLKLDS	*serine/threonine kinase 2; Serine/threonine protein kinase-2 [Homo sapiens]*
3*	Sars90(new prediction only in GZ01 strain)	GLLHRGT	NADH Dehydrogenase I Chain
4	Sars61(new prediction)	LLPLLAFL	Putative protein (Conserved across 2 organisms)
5	Sars274(Sars3a)	LLLFVTIY	Polyamine transport protein; Tpo1p *[Saccharomyces cerevisiae]*
6	Sars154(Sars3b)	QTLVLKML	K550.3.p *[Caenorhabditis elegans]*
7	Sars63(Sars6)	DDEELMEL	Elongation factor Tu *[Lactococcus lactis subsp. lactis]*
8	Sars122(Sars7a)	LIVAALVF	Putative transport transmembrane protein *[Sinorhizobium meliloti]*
		RARSVSPK	Src homology domain 3 *[Caenorhabditis elegans]*
9*	Sars78(Sars9c)	QLLAAVG	Gamma-glutamate kinase (Conserved across 8 organisms)

*: No conserved octapeptide was found. However, function has been assigned on the basis of the only highly conserved heptapeptide.

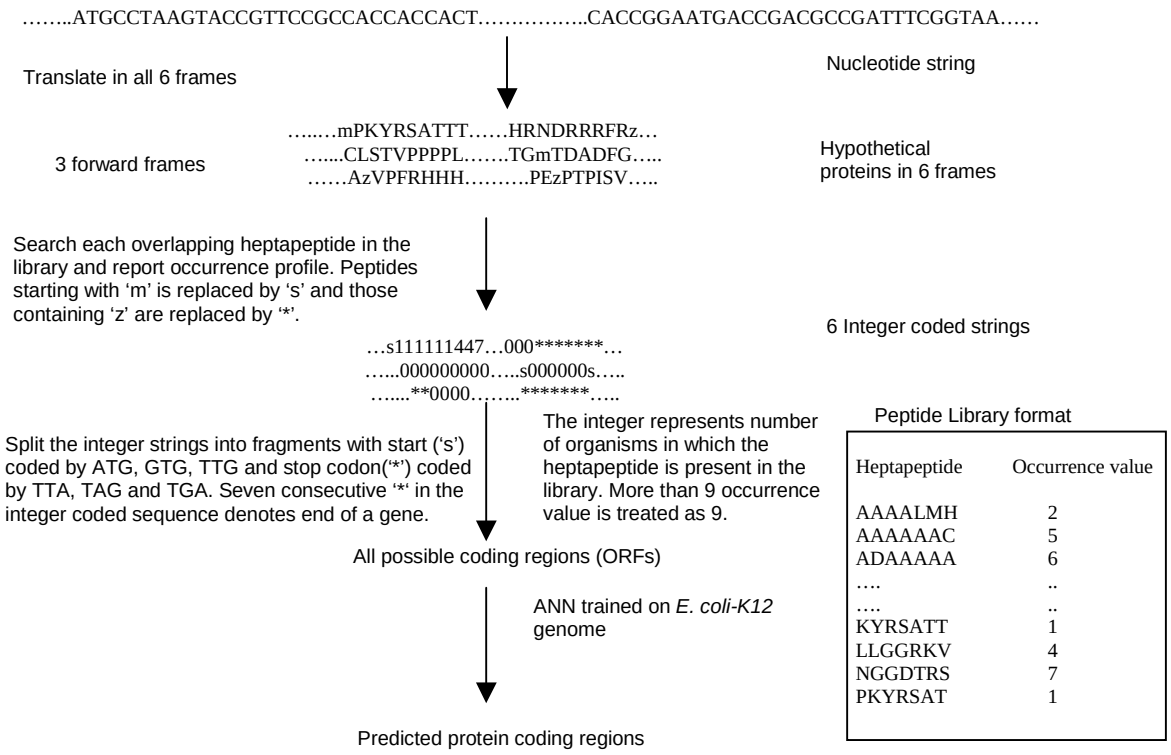

Figure 1: GeneDecipher Flow Diagram

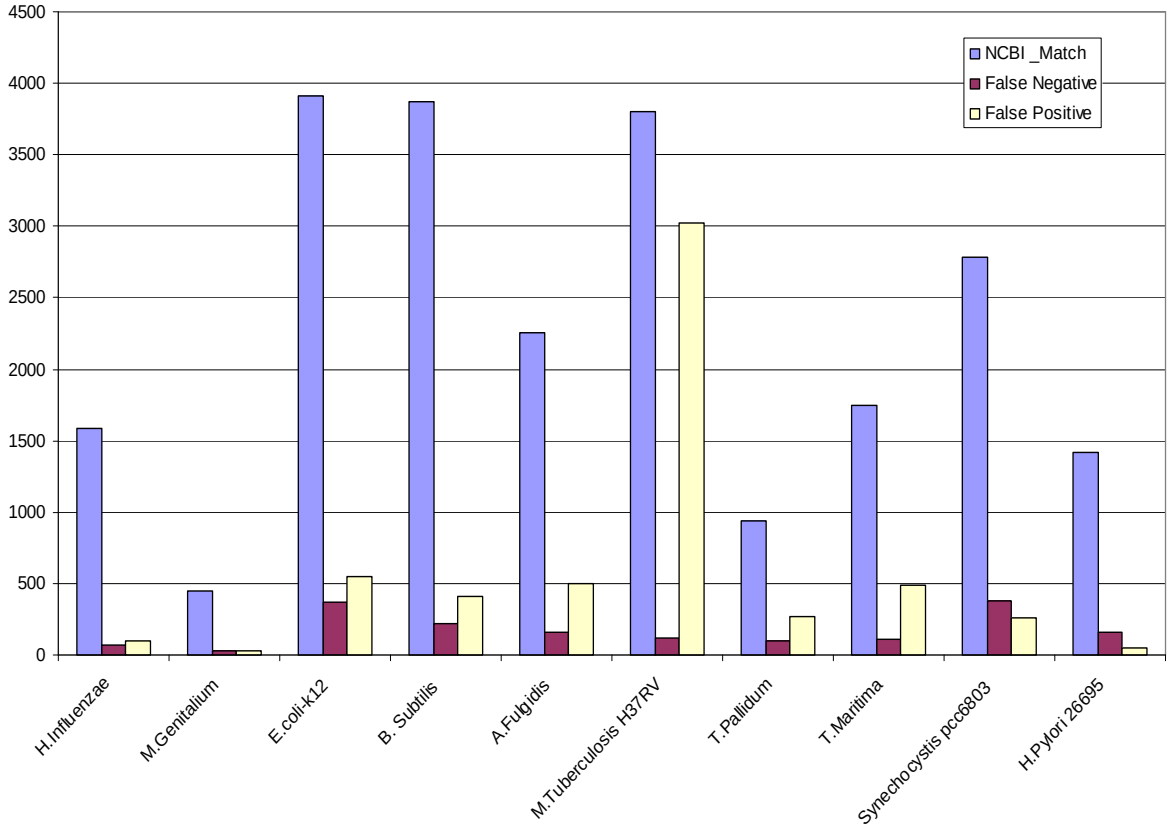

Figure2: Sensitivity and Specificity of GeneDecipher