Journal of Shanghai University (English Edition), 2003, 7(2): 118~123
Article ID: 1007-6417(2003)02-0118-06

Multiple Sequence Alignment of the M Protein in SARS-Associated and Other Known Coronaviruses

SHI Ding-Hua(史定华), *ZHOU Hui-Jie*(周晖杰), *WANG Bin-Bin*(王斌宾),
GU Yan-Hong(顾燕红), *WANG Yi-Fei*(王翼飞)
Department of Mathematics, Shanghai University, Shanghai 200436, *China*

Abstract In this paper, we report a multiple sequence alignment result on the basis of 10 amino acid sequences of the M protein, which come from different coronaviruses (4 SARS-associated and 6 others known). The alignment model was based on the profile HMM (Hidden Markov Model), and the model training was implemented through the SAHMM (Self-Adapting Hidden Markov Model) software developed by the authors.

Key words SARS (Severe Acute Respiratory Syndrome), coronavirus, M (Membrane or Matrix) protein, multiple sequence alignment, profile HMM.
MSC 2000 60J20,92C40

1 Introduction

SARS is the first newly identified serious infectious disease that human being is facing at the beginning of the 21st century. It has been primarily recognized that a variant of virus from the coronavirus family might be the candidate pathogen of SARS, as reported by WHO (World Health Organization) on April 29, 2003 (http://www.who.int/csr/sarsco untry/en).

Coronaviruses were first isolated from chickens in 1937. There are now approximately 15 species in this family. Coronavirus particles are irregularly shaped, round about 60-220 nm in diameter, with an outer envelope bearing distinctive, 'club-shaped' peplomers (round about 20nm long × 10 nm at wide distal end)[1]. This 'crown-like' appearance (Latin, *corona*) gives the family its name.

The genome size of SARS-associated coronaviruses (isolate BJ01) is 29725kb and has 11 ORFs (Open Reading Frames). The whole genome is composed of a stable region encoding an RNA-dependent RNA polymerase (composed of 2 ORFs) and a variable region representing 4 CDSs (Coding Sequences) viral structural genes (the S, E, M, N proteins) and 5 PUPs (Putative Uncharacterized Proteins)[2]. Its gene order is identical to that of other known coronaviruses.

The S (Spike) protein, the N(Nucleocapsid) protein and perhaps together with the M protein appear to be the most important candidates for the future diagnostic testing, preventing and treatment based on antibodies and vaccines, as well as exploring the immunoreactions[2]. Due to the limit of page space, we choose the M protein as an illustrated example here. The M protein with transmembrane-budding and envelope formation was predicted to be a mid-sized protein (221 acid amino residues). It was located at the nucleotide position 26379-27044 (isolate BJ01)[2].

For the M protein, by using the Blast method and the ClustalW 1.8 software (http:// www.ddbj.nig.ac.jp/E-mail/ clustalw-e.htm), the results on both the pair and the multiple sequence alignments have been respectively obtained and reported in literature. However, as far as the authors know, the multiple sequence alignment result based on the profile HMM has not been seen yet.

In this paper, we report some results about a multiple sequence alignment on the basis of 10 amino acid

Received May 22, 2003
Project supported by the National Natural Science Foundation of China (Grant No. 70171059) and the 863 Project (Grant No. 2002AA234021)
SHI Ding-Hua, Prof., E-mail: shidh2001@263.net

sequences of the M protein, which come from different coronaviruses in NCBI databases (http://www.ncbi.nlm.nih.gov). They covered 4 SARS-associated coronaviruses isolated from patients in Canada, USA, and China (Beijing, Hong Kong), and 6 others: 2 from human being (229E, Transmissible gastroenteritis), 3 from house animals (Porcine, Bovine, Turkey), and 1 from bird (Avian).

2 Model and Method

The alignment model is based on the profile HMM, and its topology as follows[3]:

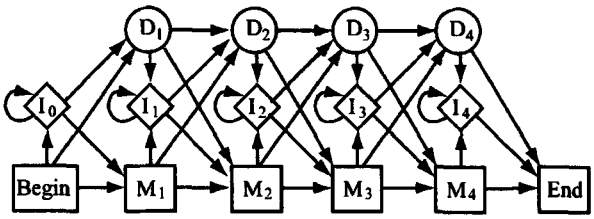

Fig.1 Topology of the profile HMM

The model training is implemented through the SAHMM software developed by the authors. The SAHMM software includes a two-stage alternative optimization method to maximize Bayesian posterior probabilities of parameters and topology for a hidden Markov model.

Let M_N denote the profile HMM with N main states, λ_N the parameter set of the profile HMM (including the state transition probabilities and the symbol emission probabilities), $O = \{O^{(w)}\}$, $w = 1,2,\cdots W$, the training sequence set, and T_w the length of the training sequence $O^{(w)}$.

The first step of two-stage alternative optimization method in the SAHMM software is parameter estimation that is to find λ_N^* as the number of main states N is fixed. By using the Bayes formula, we have

$$\lambda_N^* = \arg\max_{\lambda_N} P(\lambda_N | O, M_N) \propto \arg\max_{\lambda_N} P(O | \lambda_N, M_N) P(\lambda_N | M_N) \quad (1)$$

in which $P(O | \lambda_N, M_N)$ is the likelihood function of the training sequence set O, $P(\lambda_N | M_N)$ is the prior distribution of the parameter set λ_N. We use Bayesian Baum-Welch algorithm plus simulated annealing to estimate the parameters λ_N of the profile HMM. The Baum-Welch algorithm is a variation of the more general EM algorithm. It iterates between an expectation step (E-step) and a maximization step (M-step). The iterative process continues until some stop rule is satisfied.

The second step of two-stage alternative optimization method in the SAHMM software is the topology optimization that is to find the following M_N^*.

$$M_N^* = \arg\max_{M_N} P(M_N | O) \propto \arg\max_{M_N} P(O | M_N) P(M_N) \quad (2)$$

in which $P(M_N)$ is the prior distribution of the model topology M_N. Under the assumption of a non-information prior distribution, we have

$$P(M_N | O) \propto P(O | M_N) = \int P(O | M_N, \lambda_N) P(\lambda_N | M_N) d\lambda_N \quad (3)$$

Usually, the integral in Eq. (3) is difficult to calculate directly. Hence we use Bayesian Information Criterion (BIC)[4] to approximate it:

$$BIC = -2\log P(O | \lambda_N^*, M_N) + K_N \log W \quad (4)$$

where K_N is the number of free parameters in the profile HMM with N main states, W is the sample size, and $-\log P(O | \lambda_N^*, M_N)$ is the maximized negative log-likelihood of training sequence set O. Then the optimum topology model M_N^* is

$$M_N^* = \arg\min_{M_N} \{-2\log P(O | \lambda_N^*, M_N) + K_N \log W\} \quad (5)$$

We have proved that $P(O | \lambda_N^*, M_N)$ is a monotonously increasing function with respect to N, so the object function of (5) is a single peak function. We can use various optimum methods to solve (5), *e.g.* the golden section method.

3 Data and Results

3.1 Data

Organism	Accession	Length	Web site
SARS coronavirus BJ01	AY278488.2	221a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=30275673&dopt=GenPept
SARS coronavirus CUHK-W1	AY278554.2	221a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=30023958&dopt=GenPept
SARS coronavirus NC_004718.3	NC_004718.3	221a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list-uids=29836504&dopt=GenPept
SARS coronavirus urbani	AY278741.A	221a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=30027623&dopt=GenPept
Transmissible gastroenteritis virus	NC_002306.2	262a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=13399294&dopt=GenPept
Human coronavirus 229E	NC_002645	225a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list-uids=12175752&dopt=GenPept
Porcine-epidemic diarrhea virus	D49591	226a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=1360870&dopt=GenPept
Bovine coronavirus	AF220295.1	230a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=17529680&dopt=GenPept
Turkey coronavirus	JQ1172	230a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=77083&dopt=GenPept
Avian-infectious bronchitis virus	M95169.1	225a.a.	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=protein&list_uids=292958&dopt=GenPept

3.2 Results

The multiple sequence alignment of the M protein produced by the ClustalW (1.8) software

```
BJ01-a             ------------------------------------MADNGTITVEELKQLLEQWNLVIGFLFLAW
CUHK-b             ------------------------------------MADNGTITVEELKQLLEQWNLVIGFLFLAW
NC-c               ------------------------------------MADNGTITVEELKQLLEQWNLVIGFLFLAW
urbani-d           ------------------------------------MADNGTITVEELKQLLEQWNLVIGFLFLAW
Transmissible-e    -MKILLILACVIACACGERYCAMKSDTDLSCRNSTASDCESCFNGGDLIWHLANWNFSWSIILIVF
human-f            ------------------------------------MSNDNCTGDIVTHLKNWNFGWNVILTIF
porcine-g          ------------------------------------MSNGSIPVDEVIEHLRNWNFTWNIILTIL
Bovine-h           ----------------------------MSSVTTPAPVYTWTADEAIKFLKEWNFSLGIILLFI
Turkey-i           ----------------------------MSSVTTPAPVYTWTADEAIKFLKEWNFSLGIILLFI
Avian-j            ------------------------------------MPNETNCTLDFEQSVQLFKEYNLFITAFLLFL

BJ01-a             IMLLQFAYSNRNRFLYIIKLVFLWLLWPVTLACFVLA--AVYRIN-WVTGGIAIAMACIVFLMWLS
CUHK-b             IMLLQFAYSNRNRFLYIIKLVFLWLLWPVTLACFVLA--AVYRIN-WVTGGIAIAMACIVGLMWLS
NC-c               IMLLQFAYSNRNRFLYIIKLVFLWLLWPVTLACFVLA--AVYRIN-WVTGGIAIAMACIVGLMWLS
urbani-d           IMLLQFAYSNRNRFLYIIKLVFLWLLWPVTLACFVLA--AVYRIN-WVTGGIAIAMACIVGLMWLS
```

```
Transmissible-e ITVLQYGRPQFSWFVYGIKMLIMWLLWPVVLALTIFNAYSEYQVSRYVMFGFSIAGAIVTFVLWIM
human-f         IVILQFGHYKYSRLFYGLKMLVLWLLWPLVLALSIFDTWANWDSN-WAFVAFSFFMAVSTLVMWVM
porcine-g       LVVLQYGHYKYSVFLYGVKMAILWILWPLVLALSLFDAWASFQVN-WVFFAFSILMACITLMLWIM
Bovine-h        TVILQFGYTSRSMFVYVIKMVILWLMWPLTIILTIFN--CVYALN-NVYLGFSIVFTIVAIIMWIV
Turkey-i        TIILQFGYTSRSMSVYVIKMIILWLMWPLTIILTIFN--CVYALN-NVYLGFSIVFTIVAIIMWIV
Avian-j         TIILQYGYATRSKVIYTLKMIVLWCFWPLNIAVGVIS--CTYPPN-TGGLVAAIILTVFACLSFVG

BJ01-a          YFVASFRLFARTRSMWSFNPETNILLNVPLR-GTIVTRPLMESELVIGAVIIRGHLRMAGHSLGR-
CUHK-b          YFVASGRLGARTRSMWSFNPETNILLNVPLR-GTIVTRPLMESELVIGAVIIRGHLRMAGHSLGR-
NC-c            YFVASFRLFARTRSMWSFNPETNILLNVPLR-GTIVTRPLMESELVIGAVIIRGHLRMAGHSLGR-
urbani-d        YFVASFRLFARTRSMWSFNPETNILLNVPLR-GTIVTRPLMESELVIGAVIIRGHLRMAGHPLGR-
Transmissible-e YFVRSIQLYRRTKSWWSFNPETKAILCVSAL-GRSYVLPLEGVPTGVTLTLLSGNLYAEGFKIAGG
human-f         YFANSFRLFRRARTFWAWNPEVNAITVTTVL-GQTYYQPIQQAPTGITVTLLSGVLYVDGHRLASG
porcine-g       YFVNSIRLWRRTHSWWSFNPETDALLTTSVM-GRQVCIPVLGAPTGVTLTLLSGTLLVEGYKVATG
Bovine-h        YFVNSIRLFIRTGSWWSFNPETNNLMCIDMK-GRMYVRPIIEDYHTLTVTIIRGHLYMQGIKLGTG
Turkey-i        YFVNSIRLFIRTGSWWSFNPETNNLMCIDMK-GRMYVRPIIEDYHTLTVTIIRGHLYMQGIKLGTG
Avian-j         YWIQSIRLFKRCRSWWSFNPESNAVGSILLTNGQQCNFAIESVPMVLSPIIKNGVLYCEGQWLAK-

BJ01-a         CDIKDLPKEITVATSR-TLSYYKLGASQRVGTDSGFAAYNRYRIGNYKLNTDHAGSNDNIALLVQ--
CUHK-b         CDIKDLPKEITVATSR-TLSYYKLGASQRVGTDSGFAAYNRYRIGNYKLNTDHAGSNDNIALLVQ--
NC-c           CDIKDLPKEITVATSR-TLSYYKLGASQRVGTDSGFAAYNRYRIGNYKLNTDHAGSNDNIALLVQ--
urbani-d       CDIKDLPKEITVATSR-TLSYYKLGASQRVGTDSGFAAYNRYRIGNYKLNTDHAGSNDNIALLVQ--
Transmissible-e MNIDNLPKYVMVALPSRTIVYTLVGKKLKASSATGWAYYVKSKAGDYSTEAR-TDNLSEQEKLLHMV
human-f         VQVHNLPEYMTVAVPSTTIIYSRVGRSVNSQNSTGWVFYVRVKHGDFSAVSSPMSNMTENERLLHFF
porcine-g       VQVSQLPNFVTVAKATTTIVYGRVGRSVNASSGTGWAFYVRSKHGDYSAVSNPSAVLTDSEKVLHLV
Bovine-h        YSLSDLPAYVTVAKVS-HLLTYKRGFLDKIGDTSGFAVYVKSKVGNYRLPSTQKGSGLDTALLRNNI
Turkey-i        YSLSDLPAYVTVAKVS-HLLTYKRGFLDKIGDTSGFAVYVKSKVGNYRLPSTQKGSGMDTALLRNNI
Avian-j         CEPDHLPKDIFVCTPDRRNIYRMVQKYTGDQSGNKKRFATFVYAKQSVDTGELESVATGGSSLYT--
```

The Multiple sequence alignment of the M protein produced by the SAHMM software

```
BJ01-a          MADNGTI--- ---------T VE-E-LKQLL EQWN------ ---LVI-GFLFLAWI-----
CUHK-b          MADNGTI--- ---------T VE-E-LKQLL EQWN------ ---LVI-GFLFLAWI-----
NC-c            MADNGTI--- ---------T VE-E-LKQLL EQWN------ ---LVI-GFLFLAWI-----
urbani-d        MADNGTI--- ---------T VE-E-LKQLL EQWN------ ---LVI-GFLFLAWI-----
Transmissible-e MKILLILACV IACACGERYC AM-K-SDTDL SCRNSTASDC ESCFNG-GDLIWHLANWNFS
human-f         M-SNDNC--- ---------T GD-I--VTHL KNWNF----- ---GWN-VILTIFIV-----
porcine-g       M-SNGSI--- ---------P VD-E-VIEHL RNWNF----- ---TWN-IILTILLV-----
Bovine-h        MSSVTTPAPV YTW------T AD-E-AIKFL KEWNFS---- ---LGI--ILLFITV-----
Turkey-i        MSSVTTPAPV YTW------T AD-E-AIKFL KEWNFS---- ---LGI--ILLFITI-----
Avian-j         MPNETNC--- ---------T LDFEQSVQLF KEYN------ ---LFITAFLLFLTI-----
```

```
BJ01-a           ---------- MLLQFAYSNR NRFLYIIKLV FLWLLWPVTL A-CFVLA-AV YRI-NWVTGG
CUHK-b           ---------- MLLQFAYSNR NRFLYIIKLV FLWLLWPVTL A-CFVLA-AV YRI-NWVTGG
NC-c             ---------- MLLQFAYSNR NRFLYIIKLV FLWLLWPVTL A-CFVLA-AV YRI-NWVTGG
urbani-d         ---------- MLLQFAYSNR NRFLYIIKLV FLWLLWPVTL A-CFVLA-AV YRI-NWVTGG
Transmissible-e  WSIILIVFIT VL-QYGRPQF SWFVYGIKML IMWLLWPVVL ALTIFNAYSE YQVSRYVMFG
human-f          ---------- IL-QFGHYKY SRLFYGLKML VLWLLWPLVL ALSIFDTWAN WDS-NWAFVA
porcine-g        ---------- VL-QYGHYKY SVFLYGVKMA ILWILWPLVL ALSLFDAWAS FQV-NWVFFA
Bovine-h         ---------- IL-QFGYTSR SMFVYVIKMV ILWLMWPLTI ILTIFNC-VY ALN-NVYLGF
Turkey-i         ---------- IL-QFGYTSR SMSVYVIKMI ILWLMWPLTI ILTIFNC-VY ALN-NVYLGF
Avian-j          ---------- IL-QYGYATR SKVIYTLKMI VLWCFWPLNI A-VGVIS-CT YPP-N--TGG

BJ01-a           -IAIAMACIV G--LMWLSYF VASFRLFART RSMWSFNPET NILLNVPL-R GTIVTRPLME
CUHK-b           -IAIAMACIV G--LMWLSYF VASFRLFART RSMWSFNPET NILLNVPL-R GTIVTRPLME
NC-c             -IAIAMACIV G--LMWLSYF VASFRLFART RSMWSFNPET NILLNVPL-R GTIVTRPLME
urbani-d         -IAIAMACIV G--LMWLSYF VASFRLFART RSMWSFNPET NILLNVPL-R GTIVTRPLME
Transmissible-e  -FSIAGAIVT F--VLWIMYF VRSIQLYRRT KSWWSFNPET KAILCVSALG RSYV-LPLEG
human-f          -FSFFMAVST L--VMWVMYF ANSFRLFRRA RTFWAWNPEV NAITVTTVLG QTYY-QPIQQ
porcine-g        -FSILMACIT L--MLWIMYF VNSIRLWRRT HSWWSFNPET DALLTTSV-M GRQVCIPVLG
Bovine-h         SIVFTIVAII ----MWIVYF VNSIRLFIRT GSWWSFNPET NNLMCIDMKG RMYV-RPIIE
Turkey-i         SIVFTIVAII ----MWIVYF VNSIRLFIRT GSWWSFNPET NNLMCIDMKG RMYV-RPIIE
Avian-j          -LVAAIILTV FACLSFVGYW IQSIRLFKRC RSWWSFNPES NAVGSILLTN GQQC-NFAIE

BJ01-a           S-ELVIGAVI IRGHLRMAGH -SLGRCDIKD LPKEITVA-T SRTLSYYKLG A--SQRVGTD
CUHK-b           S-ELVIGAVI IRGHLRMAGH -SLGRCDIKD LPKEITVA-T SRTLSYYKLG A--SQRVGTD
NC-c             S-ELVIGAVI IRGHLRMAGH -SLGRCDIKD LPKEITVA-T SRTLSYYKLG A--SQRVGTD
urbani-d         S-ELVIGAVI IRGHLRMAGH -PLGRCDIKD LPKEITVA-T SRTLSYYKLG A--SQRVGTD
Transmissible-e  V-PTGVTLTL LSGNLYAEGF KIAGGMNIDN LPKYVMVALP SRTIVYTLVG K--KLKASSA
human-f          A-PTGITVTL LSGVLYVDGH RLASGVQVHN LPEYMTVAVP STTIIYSRVG R--SVNSQNS
porcine-g        A-PTGVTLTL LSGTLLVEGY KVATGVQVSQ LPNFVTVAKA TTTIVYGRVG R--SVNASSG
Bovine-h         D-YHTLTVTI IRGHLYMQGI KLGTGYSLSD LPAYVTVAKV SHLLTY-KRG F--LDKIGDT
Turkey-i         D-YHTLTVTI IRGHLYMQGI KLGTGYSLSD LPAYVTVAKV SHLLTY-KRG F--LDKIGDT
Avian-j          SVPMVLSPII KNGVLYCEGQ -WLAKCEPDH LPKDIFVCTP DRRNIYRMVQ KYTGDQSGNK

BJ01-a           SGFAAYNRYR  ---IGNYKLN TD-HAGSNDN IALL--VQ
CUHK-b           SGFAAYNRYR  ---IGNYKLN TD-HAGSNDN IALL--VQ
NC-c             SGFAAYNRYR  ---IGNYKLN TD-HAGSNDN IALL--VQ
urbani-d         SGFAAYNRYR  ---IGNYKLN TD-HAGSNDN IALL--VQ
Transmissible-e  TGWAYY-VKS  K--AGDYSTE AR-TDNLSEQ EKLLH-MV
human-f          TGWVFY-VRV  K--HGDFSAV SSPMSNMTEN ERLLH-FF
```

```
porcine-g      TGWAFY-VRS  K--HGDYSAV SNPSAVLTDS EKVLH-LV
Bovine-h       SGFAVY-VKS  K--VGNYRLP ST-QKGSGLD TALLRNNI
Turkey-i       SGFAVY-VKS  K--VGNYRLP ST-QKGSGMD TALLRNNI
Avian-j        KRFATF-VYA  KQSVDTGELE SV-ATGGS-- --SL--YT
```

References

[1] Ksiazek T G, *et al*. A novel coronavirus associated with Severe Acute Respiratory Syndrome, *The New England Journal of Medicine*, 2003, April 10.

[2] Qin E'de, *et al*. A complete sequence and comparative analysis of a SARS-associated virus (isolate BJ01), *Chinese Science Bulletin*, 2003, **48**(10): 941 - 948.

[3] Durbin S, *et al*. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, London: *Cambridge University Press*, 1998.

[4] Schwarz G. Estimating the dimension of a model, *Annuals of Statistics*, 1978, **6**: 461 - 464.

(*Executive editor SHEN Mei-Fang*)