Bioinformatics Research on the SARS Coronavirus (SARS_CoV) in China

Pei Hao[1,2], Min Chen[1], Guoqing Zhang[1,3] Weizhong He[1] and Yixue Li[1,2,*]

[1]*Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235,China;* [2]*Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai200031, China and* [3]*Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan, Hubei 430074, P.R. China*

Abstract: Severe acute respiratory syndrome (SARS) first appeared in 2002 in China, which fastly affected about 8000 patients over 29 countries and caused 774 fatalities. As its pathogen was identified as a new kind of coronavirus (SARS_CoV), its genome was quickly sequenced on several isolates. Studies on its functional genomics were performed by combinatorial application of all the available bioinformatics tools and the development of new programs. In this way, it was found that the four proteins were absolutely responsible for nosogenesis of SARS, i.e. spike (S) protein; small envelop (E) protein; membrane (M) protein; and nucleocaspid (N) protein. Molecular evolution studies have revealed that SARS must be originated from wild animals, and it was demonstrated that the major genetic variations in some critical genes, particularly the Spike gene, was essential for the transition from animal-to-human transmission to human-to-human transmission. Theoretical models, either Logistic model or SIR model, were developed to describe the transmission of SARS. The recorded difference of SARS spreading in Beijing and Hong Kong was also reasonably analyzed according to these models. The whole process of fruitful bioinformatics studies, along with other related scientific investigations have set up an unprecedented paradigm for human of how to battle against sudden-breaking and catastrophic epidemics.

Key Words: SARS, SARS_CoV, functional genomics, molecular evolution, transmission model.

INTRODUCTION

Severe acute respiratory syndrome (SARS) first emerged in November, 2002, in Guangdong Province, Southern China, and then was identified as a new coronavirus in March, 2003 [1, 2]. Later it was known that a new pathogen, a member of the Coronaviridae family of enveloped, POSITIVE-STRANDED RNA VIRUS, the SARS_CoV was origin of epidemic. Less than half year later, it had affected over 8000 patients in 29 countries with 774 fatalities. It was an unprecedented global experience in the rapidity and extent of its spreading, the magnitude of its impact on the health systems and economies, and in the effectiveness of its control [3].

After SARS broke out, Chinese government took aggressive public health measures to bring SARS under control, but at that time there were no effective drugs or vaccines against SARS, not even drastic debating on the pathogen identification. Control of this disease only relied on the rapid identification of cases and their appropriate management, including the isolation of suspect and probable cases and the management of their close contacts. By these very strict isolations it was impossible to get enough samples for laboratory experiments at the early stage of epidemic. On the other hand, there were no safe experimental equipments suitable for such virology study. Most of the life science research laboratories in China could not perform their initial studies about SARS_CoV, therefore, scientists in China had no opportunity to be the first to nail the pathogen, sequence its genomes, or describe how it sickens its victims based on the laboratory works. Although at the first stage, Chinese scientists were "defeated" [4], the researchers determined to catch-up and many institutes are now abuzz with SARS research projects after Chinese government had embraced science as a key weapon against the disease. Researches in various areas had been processed, and bioinformatics was used as one of the practical tools. Under such crucial situation, "Bioinformatics" studies were started, and now its important role is generally recognized. These studies have given a strong support to the campaigns against SARS in China, and benefited worldwide battle against this disease. Several hundreds of bioinformatics based or related scientific papers were published around the time when SARS emerged, and a many wide research fields were involved, including: the molecular biology of viruses, the pathogenic mutation, nosogenesis, the channels for the spread of the pathogeny, epitope prediction and immunity mechanisms, drug design, vaccine design and vaccination development, early diagnosis and treatment of the SARS disease, and its prevention. Generally, these studies can be grouped into three aspects of SARS_CoV research: functional genomics, which was focused on gene prediction and functional mechanisms of gene products; molecular evolution, which was mainly about genotype identify, evolution distance; and the transmission model, which could be used to predict the SARS' spread capability and help to give the evaluation to the effect of controls, related adjustment. In the following sections, we will try to summarize the latest advances of these aspects; also some of our own work is included.

*Address correspondence to this author at Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, China; Tel: +86-21-6436-3311; Fax: +86-21-6483-8882; E-mail: yxli@sibs.ac.cn

1381-6128/06 $50.00+.00

FUNCTIONAL GENOMICS STUDIES ON SARS_COV

Just after the whole genome of SARS_CoV was sequenced by scientists in Canada on 13 April, 2003[5], the complete genome sequence of another isolate of SARS, BJ01, was sequenced by Chinese scientists at BGI (Beijing Genomics Institute, CAS) [6], and 11 ORFs (open-reading frame) were identified through ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/). The physical and chemical features of proteins were predicted by using Compute PI/MW (http://us.expasy.org/tools/pi_tool.html) and also the N-Glycosylation sites were predicted by applying NetNGlyc 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/). Sooner or later, Qin *et al*. completely sequenced the genome of the GD01 isolate [7], and the isolates of BJ group (Isolates BJ01, BJ02, BJ03 and BJ04) were sequenced by BGI [8]. Additional two ORFs were identified later. All these researches carried multi-sequence alignment with the other isolates of SARS_CoV, and the mutations among different virus genomes were found. Furthermore, the multi-sequence alignment with other coronavirus genomes confirmed the early hypothesis that the SARS_CoV may be originated from animals [7, 8].

By applying several prediction methods, some groups put more efforts to predict new ORFs in order to complete SARS_CoV genome annotation. Through the comparison and evaluation of 12 ORFs by different prediction methods, Chen *et al*. selected 4 of them [9]: Heuristic models, gene identification and ORF finder, doing gene prediction of SARS_CoV genome. The ATGpr program, which is the start coden prediction software based on a linear discriminant algorithm, is usually used to calculate and estimate the probability of each ATG being the initiation codon [10]. Chen *et al*. used this program to verify possibilities of start codon and to see whether the Kozak rule [11] was followed. Taking advantage of this combinatory gene finding strategy, 21 new proteins were predicted and the appearance possibilities were analyzed by BLAST and FASTA with other sequences in nr database, and all these informations were extracted for further evaluations [12].

Developing new algorithms about mining biological data and knowledge extraction is always the most active work of bioinformatics. In searching of novel genes from SARS_CoV genome with higher accuracy and sensitivity, more efforts were also put into developing of new gene prediction algorithms (as shown in Fig. **1**). Zhang developed a new program based on the famous Z curve theory of DNA sequence [13], and named it as ZCURVE_CoV (version 1.0) [14]. This program can be used to recognize protein coding genes in coronavirus genomes with highly accurate gene start prediction, and particularly suitable for the prediction of SARS_CoV genomes. Comparing with some other available tools, this new program package has the merits of simplicity, higher accuracy, higher reliability, and quickness. Zhang's group used ZCURVE_CoV system for each of the 11 newly sequenced SARS_CoV genomes, and got good results. For example, annotating six genomes which were not annotated previously, checking and discussing of some problems on previous annotations of other five genomes. Besides the polyprotein chain ORFs 1a and 1b and the four genes coding for the major structural proteins (i.e. spike (S) protein, small envelop (E) protein, membrane (M) protein and nuleocaspid (N) protein). ZCURVE_CoV also predicted 5-6 putative proteins in length between 39 and 274 amino acids with unknown functions. Some single nucleotide mutations within these putative coding sequences have been detected and their biological implications were discussed. The improved version 2.0 of this program was applied to identify all the non-structural proteins cleaved by viral proteinases in the polyproteins [15]. The ZCURVE_CoV program can be accessed through its web server (http://tubic.tju.edu.cn/sars) and user can obtain the annotated results very quickly by pasting sequences. This program can also be freely downloaded from its website and run locally.

In order to clarify the mechanism of the nosogenesis of SARS disease, more and more researches focused on the annotation of protein functions. HMMER and BLIMPS were employed to identify the conserved domains in PFAM [16] (http://www.sanger.ac.uk/Software/Pfam/) and BLOCKS [17] (http://blocks.fhcrc.org), respectively. The S1 and S2 domains of the Spike protein were correctly predicted [9]. Other domains were analyzed and their functions were predicted by synthesizing different alignment and prediction methods. PSIPRED [18] was employed to predict corresponding protein secondary structures [12]. The transmembrane segment and the transmembrane helices of the protein M were predicted by using TMAP [19] and TMHMM2.0 [20]. The sub-cellular localization of proteins was predicted by PSORT [21]. Other programs for analyzing protein features including SAPS [22], PI and EXTCOEF [23] were widely used. These studies together with the achievements by researchers all over the world demonstrate that four proteins and the main protease (3C-like proteinase) of R protein [24] are the most important proteins associated with SARS_CoV infection, these are the S protein, the E protein, the M protein and the N protein. Based on the complete genome sequences for four isolates (BJ01-BJ04) of SARS_CoV from Beijing, China, Yang *et al*. analyzed the structure and predicted functions of the R protein by comparing with 13 other isolates of SARS_CoV and other coronavirus [24]. Rota and Marra *et al*. [5, 25] obtained very similar results by annotation. The entire ORF encodes for two major enzyme activities, RNA-dependent RNA polymerase (RdRp) and proteinase activities, namely the main proteinase 3CLP. They found the entire 15 function-related peptides, which were deemed from a complex proteolytic process of R polyprotein. Of 15 peptides, a hydrophobic domain (HOD) and a hydrophilic domain (HID) were newly identified within NSP1 of ORF1ab of SARS_CoV. Eleven highly conserved regions in RdRp and twelve cleavage sites by chymotrypsin-like protein (3CLP) had been identified as potential drug binding sites.

Since antisense RNA and RNA interference (RNAi) technologies have shown potential prospect in treating some severe diseases, it was considered as important candidate medicines against SARS. To help the design of such drugs, the prediction of probable genomic packaging signal of SARS_CoV was carried out on the comparison of genomic packaging signals of MHV and BCoV with both primary and secondary structure [26]. The primary sequence multi-alignment was carried out with Vector NTI [27] and secondary structure of genomic RNA sequences were predicted by

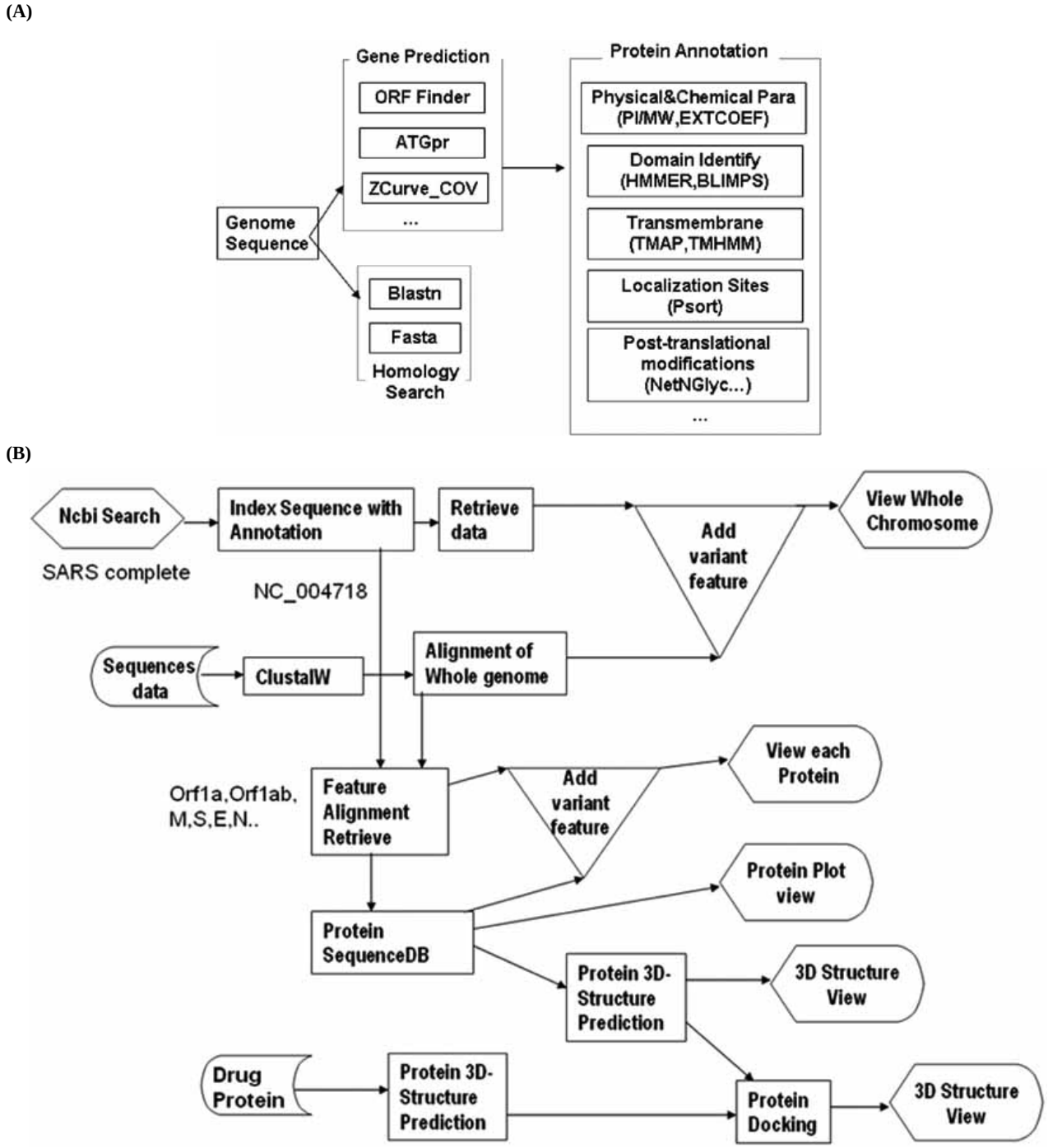

Fig. (1). (A) General pipeline for the annotation of SARS genome. (B) A detailed operating process for the gene annotation of SARS genome.

using the RNA Structure [28]. It was suggested that probable genomic packaging signal of SARS_CoV was analogous to that of MHV and BCoV, with the corresponding secondary RNA structure locating at the similar region of ORF1b, but the positions for genomic packaging signals must have suffered rounds of mutations, and inversely these mutations may influence the primary structures of the N and M proteins.

Instead of sequence based functional annotation, a number of research groups focused on the aspects of immunological mechanism of SARS_CoV infection. Among multiple steps of this approach, cell epitopes prediction is the first and the key for the preparation of antibodies of SARS_CoV. The combination of Kyte-Doolittle [29], Emini [30] and Jameson-Wolf [31] methods to predict the B-Cell epitopes is the one most frequently used. Goldkey developed by

Chinese Academy of Military Medical Sciences was employed to predict the B-Cell epitopes in the S and M proteins of SARS_CoV [32]. Another group predicted the epitopes of M and E proteins of isolates tor2 by Protean [33]. Combining at least three parameters, Lv *et al.* also considered the secondary structure of the M protein to get a more reasonable result [34]. The epitopes in M protein were predicted by these three groups are similar to one another. Of all the 9 epitopes predicted, the pairwise rate of agreement is up to 88.9%. A combination method with artificial neural network and quantitative matrix was developed and predicted six HLA-A* CTL epitope candidates in SARS_CoV N protein [35]. Integrating for analyzing the CD13 binding sites in SARS_CoV spike protein [36]. Yu *et al*. used a series of methods including comparative genomics, homology search, phylogenetic analyses, and multi-sequence alignment for sequence similarity comparison and structure comparison. They identified several domains and motifs responsible for CD13 binding at the possible binding site in the S protein of SARS_CoV [36]. This information provided indicative clues for the function study of SARS proteins and the design of anti-SARS drugs and vaccines. Further molecular modeling and molecular docking addressing the interacting features between CD13 and S protein of SARS_CoV validated the bioinformatics predictions. They found that the P585-A653 domain of CD13 was pairing with the D757-R761 motif of the SARS_CoV S protein to form the complex, so the CD13 may be a possible receptor of the SARS_CoV S protein, and its binding may result in the SARS infection. This study also provided a practically useful strategy for mapping the possible binding receptors of the proteins in a genome. The complete sequence determination of the human genome marks the start of a new era in biological sciences, with a focus shifting from sequencing to functional mechanisms of gene products. Afterwards, a number of works were focused on the genome based drug discovery. This change was also introduced to the anti-SARS research. Just after the whole genome sequence of SARS_CoV was reported [5, 6], some research groups tried to follow this change for the finding of candidate drug targets and for the design of potential drugs. Luo *et al*. [37] studied the interactions between the N protein of SARS_CoV (SARS_NP) and human cyclophilin A (hCypA) based on the work of Yu, *et al*. [38]. Nucleocapsid protein (NP) of SARS_CoV (SARS_NP) functions in enveloping the entire genomic RNA and interacts with viron structural proteins, thus playing important roles in the process of virus particle assembly and release. Protein-protein interaction analysis using bioinformatics tools indicated that SARS_NP may bind to human cyclophilin A (hCypA), and measurements by surface plasmon resonance (SPR) technology found that the equilibrium dissociation constant of this binding ranged from 6 to 160nM [37]. The probable binding sites of these two proteins were investigated by modeling the three-dimensional structure of the SARS_NP–hCypA complex, from which revealed the important interaction residue pairs between the proteins. Site-directed mutagenesis experiments were carried out for validating the binding model, whose correctness was assessed by the observed effects on the binding affinities between the proteins. The derived binding site is reliable as there is a good agreement between the SPR data and the computationally predicted mutual binding free energies for the binding of SARS-NP (or hCypA) mutants with the wild-type hCypA (or SARS-NP). This new derived SARS-NP¾hCypA interaction model might hint another possible SARS_CoV infection pathway against human cell. The whole study has set a good paradigm of how to find a possible candidate drug target starting from the whole genome information to the wet experimental test.

To study the role of individual proteins, 3D structural modeling is a practical way, for example, the S protein of SARS_CoV [31]. The 3D structure of SARS_S1B was predicted by Homology modeling based on the structure of template proteins(PDB entries 1AOF and 1NIR) [39]. The structural model was assessed by PROCHECK [32] for stereochemical quality, by Profile-3D [39] for structure-sequence compatibility, and by WHATIF program [40] for the rationality of the predicted protein-protein interactions between hAPN receptor and the S protein of SARS_CoV [36]. The experimentally measured data of S1-ACE-2 binding [39] was reasonably explained according to the constructed model.

MOLECULAR EVOLUTION OF SARS_COV

In order to escape from host defense, RNA viruses, such as coronavirus, commonly have a high rate of genetic mutation, and therefore evolve into novel viral strains. Similarly, the mutation rate of the SARS_CoV is a key factor to know how it spreads through the population. On the other hand, if it is possible to find a clear date for the last common ancestor of SARS coronavirus strains, it should be useful to understand the circumstances surrounding the emergence of the SARS pandemic and the rate at which SARS_CoV diverges. Based on this purpose, quite a few research groups performed their molecular evolution researches and tried to answer two questions: the origins and the spread of SARS_CoV.

Zhao *et al*. sequenced a large number of samples derived from patients in different time, different areas, and paid much more attention in search of the genotype characteristics of early, middle and late phase [41] (Fig. **2**). In their work, sequences that were derived directly from the patients' clinical specimens were used for statistical analysis. Sequences generated from specimens collected more than 4 weeks after disease onset were excluded. Among all of the available sequences, only 10 (GZ02, CUHK-AG01, CUHK-AG02, GZ-C, GZ-D, HZS2-A, HZS2-Fb, HSZ-A, HSZ-Bb, HSZ-Cb) of them met all the criteria. For the phylogenetic analysis, GZ02 was selected as the out-group, since it was the most divergent from all of the remaining 9 sequences. The Pamilo-Bianchi-Li model [42, 43] was used to calculate the Ks (the rates of synonymous changes) for the 6 known concatenated coding sequences (orf1a, orf1b, S, E, M, and N) of the SARS_CoV genome. In order to find the difference of evolutionary rates among different stages of the epidemic, all of samples were divided into four types: HP03E (human patient in early stage, 2002-2003), PC03 (palm civet, 2002-2003), PC04 (palm civet, 2003-2004) and HP04 (human patient, 2003-2004) for the estimation of the neutral mutation rate and the date. The MRCA illustrated that the SARS_CoV evolved in both palm civet and human [44]. It is clearly indicated that the PC03 and PC04 were not in the same primary transmission lineage. This further dem-

onstrated that SARS was from unknown origin and evolved not only in humans but also in palm civet hosts [44]. Through exhaustive calculations, they finally detected that the neutral mutation rate of the viral genome was constant, but the amino acid substitution rate of the coding sequences was slowed during the course of the epidemic. They also compared SARS_CoV from patients in different stages with animal SARS-like coronaviruses from palm civets, and found that the earliest genotypes for human SARS_CoV were very similar to the animal SARS-like coronaviruses. Major genetic variations in some critical genes, particularly the Spike gene, seemed essential for the transition from animal-to-human transmission to human-to-human transmission, which eventually caused the first severe acute respiratory syndrome outbreak of 2002-2003 [41].

As mentioned above, the genotypes of SARS_CoV originated from human and palm civet were much alike. This indicates that the origin of SARS was possibly from wild animal. To clarify this mystery, high-throughput genome sequencing based on a large number of samples was performed, and a huge amount of sequence data was generated [41]. By the calculations using bioinformatics tools, a very interesting phenomenon was detected. At the early stage of epidemic, only two main genotypes existed; in the middle stage of epidemic, more and more different genotypes emerged because of selection pressure; but in the late stage of epidemic, a dominant genotype was formed and kept at a very low evolutionary rate. This shows that the evolutionary ability can restrict lethal mutations of SARS_CoV and can also reduce mutated gene number related with forming new phenotypes.

In view of methodology, this work was the first achievement of combining genome sequencing, molecular evolution, computational biology and bioinformatics in the study of virology and epidemic. This story also shows that multidisciplinary cooperation is needed for the research on the battle against SARS_CoV.

Many other computational biology tools have been used to explore the source of SARS_CoV. Researchers used FDOD, PHYLIP for calculating the distance based on the whole genome sequence, and predicted that SARS_CoV shared the origin with Porcine epidemic diarrhea virus [45]. Some other researchers used PSI_BLAST for exploring the origin of SARS, and concluded that SARS_CoV should be

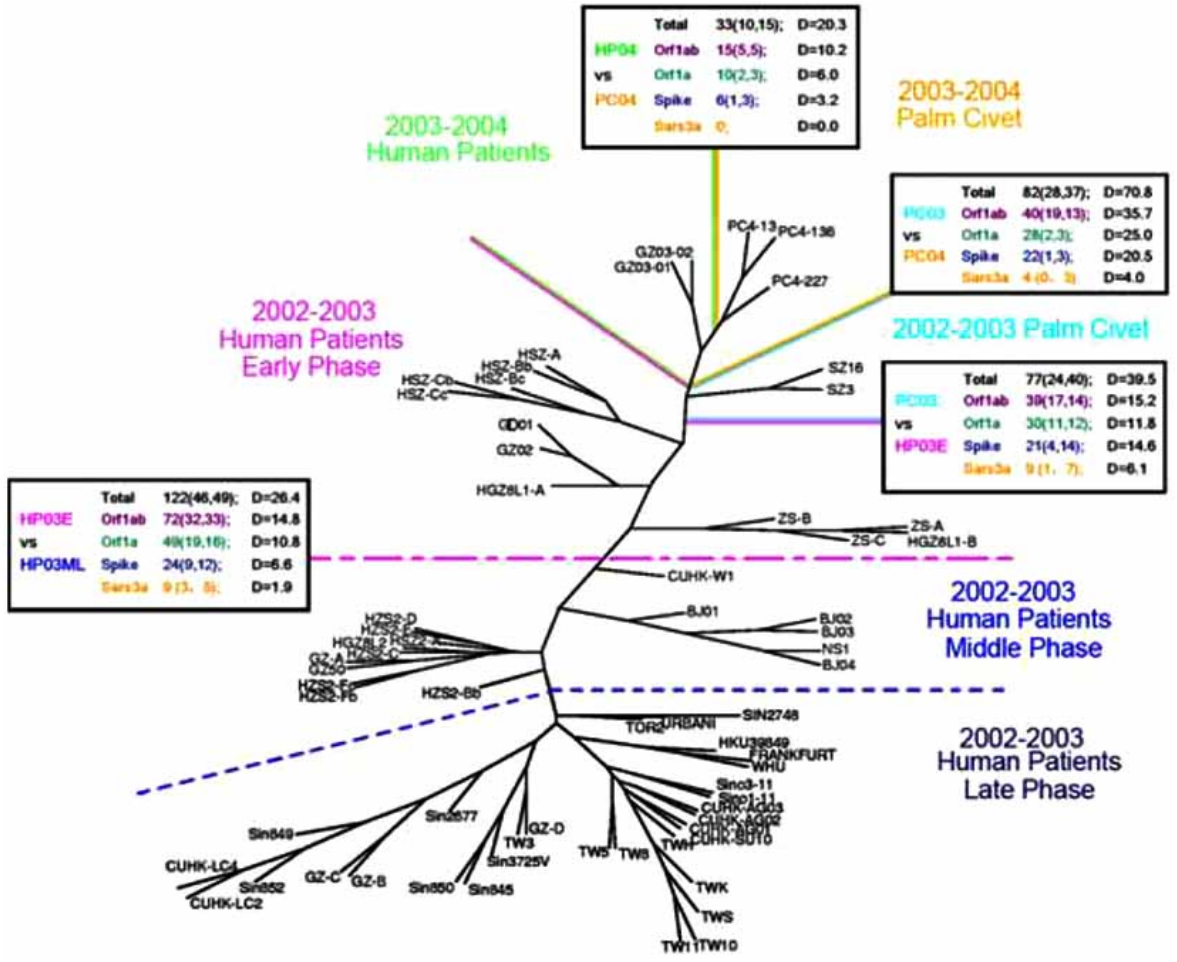

Fig. (2). Genotype clustering of SARS-CoV covering the epidemics from 2002 to 2004 illustrated by (A) an unrooted phylogenetic tree constructed with complete SNVs and deletions of 91 sequences from the human patient-derived viruses (HP) and 5 sequences from the palm civet-derived viruses (PC) and (B) a neighbor-joining (N-J) tree for the consensus nucleotide sequences of PC and early individual transmission lineages of HP. (Adopted from Song H.D. *et al*. 2005 [44]).

closer to Bovine coronavirus [12]. In addition, Lu *et al*. also proposed a new mathematical model to estimate the evolution rate of the SARS_CoV genome as well as the time of the last common ancestor of the various SARS_CoV strains [46]. Based on 6 strains with accurate dates of host death, they estimated a time of the last common ancestor and an evolutionary rate in the same range as that was reported for the HIV-1 virus [46].

On the other hand, sequence structure and features have also been studied. For example, codon preference were studied, and it was found that SARS's preference was more similar to Eukaryota than to Prokaryota, and suggested that Eukaryota such as yeast is better than others to be used to express SARS genes [47]. Dinucleotide signature of the Genome of SARS Coronavirus was analyzed and revealed that SARS coronavirus was close to the Group I coronavirus (Human coronavirus, Porcine epidemic diarrhea virus, etc.), especially transmissible guest roenteritis virus [48].

TRANSMISSION MODELS OF SARS

Many models were put forward to describe the transmission of SARS, Most of them can be grouped into either definite model or random model. Logistic model is the simplest definite model. Zheng *et al*. built the following Logistic model [49, 50] to describe the SARS:

$$\frac{dN}{dt} = rN - kN^2 = N(r - kN) \quad (1)$$

$$N_{t=t_0} = N_0 \quad (2)$$

Where N is the number of accumulated infected patients, t is the day, k is death rate, which not only depends on the SARS virus itself, but also on the effect of prevention and control. It has a direct influence on the epidemic, disease-lasting time, disease-peak appearance time and accumulated patients' number. R means disease incidence rate, and larger r means that the disease-development varies faster, and increasing rapidly from initial period to peak time. The larger k means that the patients' number is smaller at the peak time.

SIR Model is a random model put forward by Chen and Huang *et al*. [51-53].

$$\Pr\{dN(t) = 1, dR(t) = 0 \mid F_t\} = \beta \bar{S}(t) I(t) dt + o(dt) \quad (3)$$

$$\Pr\{dN(t) = 0, dR(t) = 1 \mid F_t\} = \gamma I(t) dt + o(dt) \quad (4)$$

$$\Pr\{dN(t) = 0, dR(t) = 0 \mid F_t\} = 1 - \beta \bar{S}(t) I(t) dt - \gamma I(t) + o(dt) \quad (5)$$

In above equations, N(t) is the accumulated number of infected patients at time t, I(t) is the number of infectable persons, S(t) is the number of susceptible persons, R(t) is cured patients and dead patients, N(t) is total number. $S(t) = S(t)/n$, is the proportion of susceptible number to total number.

Equation (3) describes the probability of an additional infector in a very short time period after time t. It shows that this probability has positive correlation to I(t) and $\bar{S}(t)$. β is the probability of healthy person getting sick after contacting with an infectable person, thus it is the transmission rate from susceptible to infected. γ in equation (4) is the reciprocal of average infecting time, which is the possibility from infected state to the dead or the cured, thus considered as the removal rate. μ ($\mu=\beta/\gamma$) is the basic reproduction rate, and can reflect the ability of how many people inducted as patients can be inducted during the induction period. When μ is smaller than 1, inducted population will be small and the infectious disease will disappear by itself. However, when μ is larger than 1, the infectious disease will spread continuously and induct more persons.

Chen *et al*. [51] compared the SARS data about the number of infected and latent people in Beijing with Hong Kong. It was found that Hong Kong's β value was smaller than Beijing, and the decreasing speed for μ value to 1 was faster than that of Beijing,

It should be pointed out that the two models are well fitted from the real epidemic kinetics of SARS. The logistic model is more powerful for the description of the guard effect, while the SIR model is better for the description of the SARS' spread capability without any disturbance. With the help of these simple mathematical models, the government and the related organizations can forecast the number of people that will be infected by SARS and how long SARS will continue. What's more, with the evaluation to the effect of controls, related adjustment can be applied.

CONCLUDING REMARKS

Three years ago, the initial cases of SARS appeared in the Guangdong Province in the South China. Afterwards, this terrible epidemic started to spread around the world. It is the first wild spreading and easy-infected disease of this new century, especially of its unusually high morbidity and mortality rates. It had taken full advantage of the opportunities provided by a world of international travel. At that time, China had no effective drugs or usable vaccines against SARS, so the control of this terrible disease had to rely on the rapid identification of susceptible cases and effective management of the discovered cases, including the isolation of the suspects with their close contacts. These measures prevented imported cases from spreading the disease to others, but also brought disadvantage for the scientific research at laboratory level.

As soon as the genome sequence of SARS_CoV was released from NCBI, many research groups in China took tools in bioinformatics in combination with a variety of other practical ways to battle against SARS in trying to first find the origin of SARS_CoV, and then studied the function of its genes, protein-protein interactions, possible epitopes, protein structures and the search of drug target candidates. Most of the bioinformatics researches as described above are fruitful and meaningful. It is no doubt that "Bioinformatics" research works helped China to gain time for further biomedical experiments against SARS, and in general it also provided a useful scientific paradigm on how bioinformatics research can work together with other areas in the face of sudden-breaking and catastrophic epidemics.

Recently, the specific science meeting held in Washington D. C of USA heard that the strain of SARS_CoV, which was jumping readily between humans, may initially exist in

laboratory samples. This would be an unhappy accident. There may be a fresh mutation of the virus in animal host for this re-emerging. And even if it did, it could be quickly contained (http://news.bbc.co.uk/1/hi/health/4280253.stm). As a whole, all the researches on SARS_CoV are certainly helpful in the future for the worldwide battle against other spreading human diseases caused by unknown coronaviruses.

REFERENCES

References 54-56 are related articles recently published in Current Pharmaceutical Design.

[1] Drosten C, Preiser W, Gunther S, Schmitz H, Doerr HW. Severe acute respiratory syndrome: identification of the etiological agent. Trends Mol Med 2003; 9: 325-327.

[2] Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S. A novel coronavirus associated with severe acute respiratory syndrome. N Eng J Med 2003; 348: 1953-1966.

[3] Poon LL, Guan Y, Nicholls JM, Yuen KY, Peiris JS. The aetiology, origins, and diagnosis of severe acute respiratory syndrome. Lancet Infect Dis 2004; 4: 663-671.

[4] Enserink M. SARS in China. China's missed chance. Science 2003; 301: 294-296.

[5] Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS. The Genome sequence of the SARS-associated coronavirus. Science 2003; 300: 1399-1404.

[6] Qin E, He X, Tian W, Liu Y, Li W, Wen J. A complete sequence and comparative analysis of a SARS-associated Virus Isolate BJ01. Chinese Sci Bull 2003; 48: 941-948.

[7] Qin E, He X, Tian W, Liu Y, Li W, Wen J. A genome sequence of novel SARS-CoV isolates: the genotype, GD-Ins29, leads to a hypothesis of viral transmission in South China. Genomics Proteomics Bioinformatics 2003; 1: 101-107.

[8] Bi S, Qin E, Xu Z, Li W, Wang J, Hu Y. Complete genome sequences of the SARS-CoV: the BJ Group (Isolates BJ01-BJ04). Genomics Proteomics Bioinformatics 2003; 1: 180-192.

[9] Chen TG, Wu SF, Wan P, Du CJ, Li JQ, Li D. Gene prediction and function research of SARS-CoV(BJ01). Yi Chuan Xue Bao 2003; 30: 773-780.

[10] Salamov AA, Nishikawa T, Swindells MB. Assessing protein coding region integrity in cDNA sequencing projects. Bioinformatics 1998: 14: 384-390.

[11] Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 1987; 15: 8125-8148.

[12] Liu SQ, Guo T, Ji XL, Sun ZR. Bioinformatical study on the proteomics and evolution of SARS_CoV. Chin Sci Bull 2003; 48: 1277-1211.

[13] Guo FB, Ou HY, Zhang CT. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res 2003; 31: 1780-1789.

[14] Chen LL, Ou HY, Zhang R, Zhang CT. ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. Biochem Biophys Res Commun 2003; 307: 382-388.

[15] Gao F, Ou HY, Chen LL, Zheng WX, Zhang CT. Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. FEBS Lett 2003; 553: 451-456.

[16] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S. The Pfam protein families database. Nucleic Acids Res 2004; 32: D138-D141.

[17] Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res 2000; 28: 228-230.

[18] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000; 16: 404-405.

[19] Persson B, Argos P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 1994; 237: 182-192.

[20] Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Sys Mol Biol 1998; 6: 175-182.

[21] Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999; 24: 34-36.

[22] Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S. Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 1992; 89: 2002-2006.

[23] Subramaniam S. The Biology Workbench--a seamless database and analysis environment for the biologist. Proteins 1998; 32: 1-2.

[24] Xu ZY, Zhang HQ, Tian XJ, Ji J, Yang HM. The R Protein of SARS_CoV: Analysis of Structure and Function Based on Four Complete Genome Sequences of Isolates BJ01-BJ04. Genomics Proteomics Bioinformatics 2003; 1: 156-165.

[25] Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 2003; 300: 1394-1399.

[26] Qin L, Xiong B, Luo C, Guo ZM, Hao P, Su J. Identification of probable genomic packaging signal sequence from SARS-CoV genome by bioinformatics analysis. Acta Pharm Sin 2003; 24: 489-496.

[27] Shiang R. Vector NTI Suite v5.5.1. Biotech Software & Internet Report 2000; I: 214-219.

[28] Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999; 288: 911-940.

[29] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 198; 157: 105-132.

[30] Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol 1985; 55: 836-839.

[31] Jameson BA, Wolf H. The antigenic index: a novel algorithm for predicting antigenic determinants. Comput Appl Biosci 1988; 4: 181-186.

[32] Li WJ, Fan M. Prediction of the epitopes of SARS_CoV. Med J Chin PLA 2003; 28: 9-12.

[33] He F, Zhao W, Long BG, Jian LF. Comparison of E,M genome sequence of SARS_CoV and prediction of related B cell epitopes. Prog Microbiol Immu 2004; 32: 10-14.

[34] Lv YB,Wu YZ. Prediction of the Secondary Structure and B Cell Epitopes for the M Protein of SARS Coronavirus. China Biotech 2003; 23: 41-45.

[35] Wang Q, Wu YZ. Prediction of HLA-A* 0201-restricted CTL epitopes in SARS_CoV N protein. Immunol J 2004; 20: 86-85.

[36] Yu XJ, Luo C, Lin JC, Hao P, He YY, Guo ZM. Putative hAPN receptor binding sites in SARS_CoV spike protein. Acta Pharm Sin 2003; 24: 481-488.

[37] Luo C, Luo H, Zheng S, Gui C, Yue L, Yu C. Nucleocapsid protein of SARS coronavirus tightly binds to human cyclophilin A. Biochem Biophys Res Commun 2004; 321: 557-565.

[38] Yu XJ, Lin JC, Chen M, Shi TL, Li YX. A Novel Domain-based Method to Predict the Functional Classes of Proteins. Chinese Sci Bull 2004; 49: 2379-2384.

[39] Yu C, Gui C, Luo H, Chen L, Zhang L,Yu H. Folding of the SARS coronavirus spike glycoprotein immunological fragment (SARS_S1b): thermodynamic and kinetic investigation correlating with three-dimensional structural modeling. Biochemistry 2005; 44: 1453-1463.

[40] McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994; 238: 777-793.

[41] Chinese Sars Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 2004; 303: 1666-1669.

[42] Pamilo P, Bianchi NO. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 1993; 10: 271-281.

[43] Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 1993; 36: 96-99.

[44] Song HD, Tu CC, Zhang GW, Wang SY, Zheng K, Lei LC. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci USA 2005; 102: 2430-2435.

[45] Zheng Q. The analysis of SARS_CoV based on comparative genome. Chinese Sci Bull 2003; 48: 12-12.

[46] Lu H, Zhao Y, Zhang J, Wang Y, Li W, Zhu X. Date of origin of the SARS coronavirus strains. BMC Infect Dis 2004; 4: 3-3.
[47] Wang Y, Ma WL, Zheng WL. Analysis of SARS Coronavirus' Codon Preference. Life Sci Res (Seppl) 2003; 7: 219-223.
[48] Wang Q, Li G, Li BJ. Dinucleotide Signature Analysis of the Genome of SARS Coronavirus. Chinese J Vir 2004; 20: 79-80.
[49] Zheng L, Xi C, Hu T, Long XZ. Infectious Kinetics of SARS Epidemic. Prog Biochem Biophys 2004; 31: 167-171.
[50] Huang DS, Guan P, Zhou BS. Fitness of morbidity and discussion of epidemic characteristics of SARS based on logistic model. Chinese J Public Health 2003; 19: 1-2.
[51] Chen QZ. Appl ication of SIR Model in forecasting and analyzing for SARS. J Peking Univ (Health Sci) 2003; 35: 75-80.
[52] Huang DS, Guan P, Zhou BS. Research on fitting of SIR model on prevalence of SARS in Beijing city. Chinese J Dis Contl Prev 2004; 8: 398-401.
[53] Chen Q. Application of SIR model in forecasting and analyzing for SARS. Beijing Da Xue Xue Bao 2003; 35(Suppl): 75-80.
[54] Alper O, Bowden ET. Novel insights into c-Src. Curr Pharm Des 2005; 11(9): 1119-30.
[55] Altmeyer R. Virus attachment and entry offer numerous targets for antiviral therapy. Curr Pharm Des 2004; 10(30): 3701-12.
[56] Kantardjieff K, Rupp B. Structural bioinformatic approaches to the discovery of new antimycobacterial drugs. Curr Pharm Des 2004; 10(26): 3195-211.