KO-01 Protein modeling database for novel influenza virus (A/H1N1)

Genki Terashi[*1], Kazuhiko Kanou[1], Yuuki Nakamura[1], Daisuke Takaya[2], Takehisa Matsumoto[2], Mayuko Takeda-Shitaka[1,2] and Hideaki Umeyama[1,2]

1) School of Pharmacy, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan
2) RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Introduction

NCBI (National Center for Biotechnology Information) of the United States of America is established in 1988 as a national resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. All submitted influenza sequences are available in GenBank as soon as they are processed. The swine-origin influenza A/H1N1 virus sequences isolated during human influenza outbreak of 2009 have been listed on NCBI Influenza Virus Resource site (http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html) from April 2009 [1]. As shown in this site, the result of RPS-BLAST against Protein Data Bank (PDB) database, and a summary of amino acid differences in proteins of these viruses are available at Riken National Institute of Japan, which is formally called RIKEN reorganized in 2003 as an independent administrative institution under the Ministry of Education, Culture, Sports, Science and Technology in Japan. The novel influenza (swine-origin influenza A/H1N1) protein structure model database (http://mammalia.gsc.riken.jp/swine_influenza/swine_influenza.html) [2] of RIKEN mentioned above is one part of RIKEN FAMSBASE (http://famshelp.gsc.riken.jp/famsbase/). RIKEN FAMSBASE which is protein structure model database contains more than 6,000,000 protein structure models together with the useful information as a main part. Accordingly, the above novel influenza protein structure model database contains the protein structure modeling data of ten amino acid sequences coded in the eight RNA segments in changing swine-origin influenza A/H1N1 virus genomes isolated during human influenza outbreak of 2009, and it is improved every day.

That all submitted influenza sequences are available in GenBank as soon as they are processed is corresponding to that the swine-origin influenza A/H1N1 virus sequences are listed on this page every day. That is, it is shown that, in the almost cases, the swine-origin influenza A/H1N1 virus sequences submitted by many researchers in the world are resistant to well known anti-influenza drugs such as zanamivir (relenza) and oseltamivir phosphate (tamiflu) or not. In Science journal, by analyzing the outbreak in Mexico, early data on international spread, and viral genetic diversity, they made an early assessment of transmissibility and severity. Their estimates suggested that 23,000 (range 6,000-32,000) individuals had been infected in Mexico by late April in 2009, giving an estimated case fatality ratio (CFR) of 0.4% (range 0.3% to 1.5%) based on confirmed and suspect deaths reported to that time [1]. Moreover, they are saying that, while substantial uncertainty remains, clinical severity appears less than that seen in 1918 but comparable with that seen in 1957. We have created the homology models for protein sequences published and changed every day by using our originally developed methods, following to the site of http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html in order to add the protein structure

terashig@pharm.kitasato-u.ac.jp

models of amino acid sequences newly found and experimentally analyzed. Ten kinds of modeling proteins of (1) PB2 gene for polymerase β2, segment 1, genomic RNA, (2) PB1 gene polymerase β1, segment 2, genomic RNA, (3) PA gene polymerase α, segment 3, genomic RNA, (4) HA gene for hemagglutinin, segment 4, genomic RNA, (5) NP gene for nucleoprotein, segment 5, genomic RNA, (6) NA gene for neuraminidase, segment 6, genomic RNA, (7) (8) M1 gene for matrix protein 1 and M2 gene for matrix protein 2, segment 7, genomic RNA, and (9) (10) NS1 gene for nonstructural protein 1 and NS2 gene for nonstructural protein 2, segment 8, genomic RNA. Three proteins of polymerase β2, polymerase α and neuraminidase were modeled as the protein including the low molecular weight compound, the protein-protein complex molecules and the protein including ligand such as zanamivir (relenza) or oseltamivir phosphate (tamiflu) of low molecular weight compound, respectively. Other seven proteins except for the above three proteins are modeled as the isolated protein. In the case of neuraminidase, it is easily estimated from the protein structure model that the mutation sites of amino acids newly published proteins are near the zanamivir (relenza) or oseltamivir phosphate (tamiflu) binding site or not. Moreover, in the cases of polymerase β2 and polymerase α, finding of the effective anti-virus drugs of low molecular weight compounds in the docking process with the modeled protein of each mutated influenza virus genome may be possible. As the protein structure model database of our site [2] are improved every day, the expanding areas of the same swine influenza virus genome are correlated with the danger areas of infection from the visualization of the zanamivir (relenza) or oseltamivir phosphate (tamiflu) docking with the neuraminidase enzyme.

In the protein modeling research area, thus, this protein structure model database improved every day may be an example in the first time to be useful for the Public Health.

Methods

Protein structure modeling

Target sequences for protein structure modeling were obtained from the NCBI Influenza Virus Resource site [1]. In the NCBI Influenza Virus Resource site, strains of the swine-origin influenza A/H1N1 virus are listed, and, for each strain, ten protein sequences encoded in eight RNA segments are available. Protein structure modeling for each protein sequence was performed as following procedure, and it is improved every day in relation to the increasing of target protein sequences listed on the NCBI Influenza Virus Resource site.

First, sequence alignments between the target and template proteins were obtained from executing RPS-BLAST [4] against the PDB database. The PDB database of April 2009 was used in the first version of the model database, and the PDB database had been updated twice until now, in June and September 2009, in relation to the publication of the crystal structure of RNA polymerase PB1-PB2 subunits (PDB code: 3a1g and 2ztt) and the hemagglutinin structure including some ligands (PDB code: 3hto、3htp、3htq and 3htt), respectively.

Second, sequence alignments with E-value < 0.001 were adopted, and three dimensional (3D) structure of the target protein was constructed with the Full Automatic protein Modeling System (FAMS) program [5] based on each sequence alignment. In the modeling process, FAMS moves the main chain and the side-chain atoms of the target protein alternatively in maintaining the conformational space between the model and the template 3D structure, and performs the

terashig@pharm.kitasato-u.ac.jp

conformational search iteratively as close as possible to the native structure in the packing state of the main chain and the side-chains. In the Critical Assessment of Fully Automated Structure Prediction (CAFASP-2) (2000) experiment, which is one category of CASP4 experiment, and CAFASP-3 (2002) experiment, which is one category of CASP5 experiment, FAMS was recognized as the good software for homology modeling [6,7].

Two proteins of polymerase β2 and neuraminidase were modeled as the protein including the low molecular weight compounds. M7GTP and zanamivir or oseltamivir phosphate are included in polymerase β2 and neuraminidase, respectively. Furthermore, three protein of polymerase β2, polymerase β1 and polymerase α were modeled as protein-protein complex molecule, i.e., complex of polymerase β2 and polymerase β1 and complex of polymerase β2 and polymerase α. These complex molecules were constructed using the FAMS Ligand&Complex program [8].

Results and Discussion

Table 1 shows the number of the protein sequences, the number of the protein sequence with the elimination of redundancy, the number of modeled protein sequences and number of models, for 10 proteins coded in eight RNA segments of influenza virus genomes as of September 12 2009. The number of protein sequences with redundancy elimination of PB2, HA and NA were 136, 334 and 226, respectively, and the ratios, 24, 31 and 22%, respectively, against number of protein sequences are relatively larger than those of the other proteins. This indicates that, in the PB2, HA and NA genes, the mutation arises frequently.

Protein	(1) Number of protein sequences	(2) Number of protein sequences with redundancy elimination *1	(3) Number of modeled sequences*2	(4) Number of models*3
PB2	566	136 (24.0%)	136	895
PB1	559	91 (16.3%)	90	129
PA	546	100 (18.3%)	100	397
HA	1076	334 (31.0%)	334	6084
NP	666	97 (14.6%)	95	190
NA	1045	226 (21.6%)	225	2923
M1	898	97 (10.8%)	86	175
M2	881	66 (7.5%)	62	263
NS1	620	69 (11.1%)	69	559
NS2	621	45 (7.2%)	45	177
Total	7478	1261 (16.9%)	1242	11792

Table 1. Statistics of novel influenza model database as of September 12 2009
*1 Ratio of (2) for (1) is in parenthesis.
*2 The number of sequences which have at least one model.
*3 Total number of constructed models for the sequences with redundancy elimination.

terashig@pharm.kitasato-u.ac.jp

Neuraminidase is the target protein of two anti-virus drugs, zanamivir and oseltamivir, which are generally used all over the world. Therefore, influenza virus having the mutation of amino acid residues positioned near the binding site of anti-virus drugs may be tolerized for anti-virus drugs. For some influenza virus strains, actually, it has been published in the website of NCBI that His274 near the binding site to neuraminidase is substituted to tyrosine residue (H274Y), and the strains may be changed to the virus resistant to oseltamivir. The neuraminidase 3D structures with H274Y have been also reported in our website [2]. It is noticed that the changed 3D structure of H274Y neuraminidase are immediately published in our website. This means that our modeling website may warn the appearance of oseltamivir resistant strain, and people living in the appearance region become very careful to be infected by the oseltamivir resistant virus.

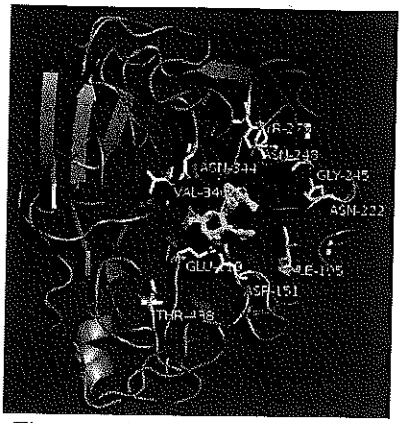

Fig.1 An example of neramnidase model in our database. Yellow colored sticks are residues positioned near the binding site of anti-virus drugs

Conclusion

We have created the protein structure model database for novel influenza A/H1N1 virus genome. The mutation rate of the influenza virus is very high, and mutated protein sequences are published every day in the NCBI Influenza Virus Resource site. Our modeling database [2] provides 3D structures of influenza virus protein and summary of amino acid differences about one day behind the publication of new amino acid sequences. Since some proteins such as polymerase β2 and neuraminidase were modeled as the protein including the low molecular weight compounds, these structure models may be useful for developing new anti-virus drugs using structure based drug-design (SBDD). Moreover, in the case of neuraminidase, it is easily estimated from the protein structure model that the mutation sites of amino acids newly published proteins are near the zanamivir (relenza) or oseltamivir phosphate (tamiflu) binding site or not. Therefore, our site of model database may be able to warn for the appearance of oseltamivir or zanamivir resistant viruses. Thus, this protein structure model database improved every day may be an example in the first time to be useful for the Public Health.

References

1. NCBI Influenza Virus Resource site: http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html
2. http://mammalia.gsc.riken.jp/swine_influenza/swine_influenza.html
3. Fraser C. et. al. Pandemic potential of a strain of influenza A (H1N1): early findings., *Science*, **324(5934)**,1557–1561 (2009).
4. Altschul S. F., Madden T. L., Schäffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J., *Nucleic Acids Res.*, **25**, 3389–3402 (1997).
5. Ogata K, & Umeyama H., J Mol Graph Model., **18(3)**, 258–272, 305–306 (2000).
6. Fischer D., Elofsson A., Rychlewski L., Pazos F., Valencia A., Rost B., Ortiz A. R., Dunbrack R. L. Jr., *Proteins*, **Suppl 5**, 171–183, (2001).
7. Fischer D., Rychlewski L., Dunbrack R. L. Jr, Ortiz A. R., Elofsson A., *Proteins*, **53 Suppl 6**, 503–516 (2003).
8. Takeda-Shitaka M., Terashi G., Chiba C., Takaya D., Umeyama H., *Med. Chem.*, **2**, 191–201 (2006).

terashig@pharm.kitasato-u.ac.jp