WEBSITE DEVELOPMENT FOR THE ADVANCEMENT OF INFLUENZA PRIMER AND PROBE DESIGN

Influenza Primer Design Resource (IPDR)

Michael E. Bose, Jiang Fan, MD, Andrew Patzer, Jack Littrell, Andrea J. Kraft, Kelly J. Henrickson, MD

Medical College of Wisconsin

Children's Research Institute

INTRODUCTION

The attention of the world and its scientific community has intensified on all aspects of influenza since the 1997 Hong Kong avian influenza epizootic. This focus has increased even more in the last few years because of the return of H5N1 in 2003. Significant resources have been invested by the United States (NIH, CDC, USDA, etc.) and a large number of countries in all areas of influenza biology but especially in the areas of vaccines, therapeutics, and diagnostics. A result of these efforts has been a significant increase in the amount of genomic data for influenza (TIGR, etc.), which is now greater than 44,000 nucleotide sequences and growing by hundreds weekly. Each year considerable resources are being spent on bioinformatics for primer and probe design by researchers or funding agencies trying to improve influenza diagnostics (1-4). Even though the NIH has funded improved bioinformatics for a number of agents including influenza (www.biohealthbase.org), the tools available are still limited, involve multiple manipulations, are focused on large oligonucleotide arrays, and do not greatly facilitate primer and probe design (5).

To solve this problem we have created an automatically updating database, containing all accessible nucleotide sequences of influenza A, B, and C, which will be integrated into a website (irwin.hmgc.mcw.edu/IPDR) that allows all 44,000+ sequences to be searched with a variety of criteria including: gene segment, year, species, geographic location, and subtype. The website quickly aligns the sequences using MAFFT and displays the consensus sequence with the percent conservation at each position and the percentage of sequences contributing to every base pair decision (6,7). Additionally, the consensus sequences are BLASTed against a database of published primers and probes and displayed with the output aligned below the sequence (8). Each consensus sequence is also analyzed by Primer3 to identify potential primers and probes (9). We believe this website will significantly advance the development of useful diagnostic assays for human and animal influenza, decrease enormous expenditures in resources around the world, and allow for rapid response as new influenza strains emerge to cause both yearly epidemics and the next pandemic.

WEBSITE DEVELOPMENT

The Database

An Oracle database of all influenza nucleotide sequences was created and updated using a perl script to download and extract the sequence information from LANL (www.flu.lanl.gov/) and NCBI (www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) on a weekly basis. The database will be set to be automatically updated every week with additional (and all) sources of genetic data publicly available.

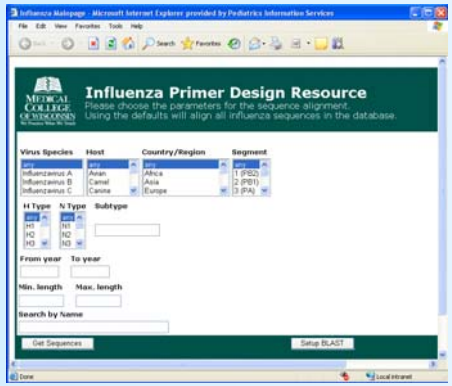
Database Statistics (1/17/2007)

Total Number of Sequences = 44186

Species		Gene Segment	
Influenza A	39523	1 PB1	4190
Influenza B	3812	2 PB2	4170
Influenza C	695	3 PA	4120
		4 HA	11272
Region		5 NP	4757
Asia	16856	6 NA	5831
Africa	420	7 MP	5036
Europe	3746	8 NS	4809
North America	15801		
Oceania	6496	**Year**	
South America	440	2000-Present	26647
		Before 2000	17530

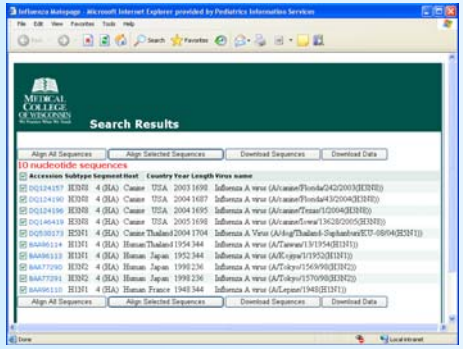
The Web Interface

The database is linked to a web interface that allows the user to search the database for sequences based on a variety of criteria including: host, segment, subtype, country/region, year, length, and name. The main page of the website serves two purposes. It allows the user to either search the database for sequences to align or for sequences to run a BLAST against. The main reason for including the BLAST option is to allow the user to determine whether their primers/probes are specific for subgroups of influenza, such as the hemagglutinin gene from the H1 subtypes.



The Search Results

The results of the search are returned to the user in a list format so they can either manually select which sequences they want to perform the alignment on or choose to have the alignment performed on all of the sequences. The user also has the option to download the sequences that were returned by the search in FASTA format. The user may also download a tab-delimited file that includes the accession number, host, segment, subtype, country, year, length, and name for each sequence returned by the search.

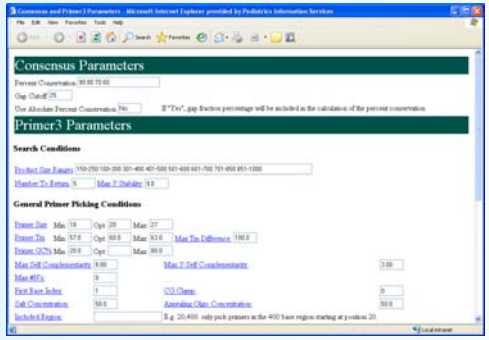


Consensus and Primer3 Parameter Selection

After the user selects the sequences that they want to align they are taken to the Consensus and Primer3 Parameter screen, on this screen they have the option to adjust multiple parameters that are used for generating their output. For the consensus parameters the user has the option to input multiple conservation percentages to use as a cutoff, for which a consensus sequence will be generated for each one. The user can also adjust the "gap cutoff", which is the minimum percentage of sequences with a nucleotide in the aligned position required for it to not be considered a gap in the consensus sequence (e.g., gap cutoff=10, then to show in the consensus sequence a nucleotide at any one position >=10% of aligned sequences must have a nucleotide at that same position). A third parameter is the Use Absolute Percent Conservation parameter. By default, when calculating the percentages at each position for the consensus sequence the program uses the following formula:

$$\frac{\text{\# of Sequences with Nucleotide}}{\text{Total \# of Sequences - \# of Sequences with Gap}}$$

When using the Absolute Percent Conservation, the number (#) of Sequences with Gap are not used in the calculation. For Primer3 the user is able to adjust a variety of standard Primer3 parameters. Once the parameters are selected the analysis is performed.

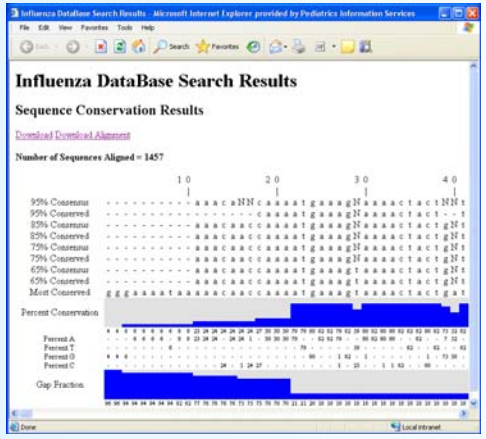


The Analysis Step

The analysis that is performed involves several steps. First, the sequences are aligned using MAFFT under the default conditions. MAFFT was selected for its ability to align a large number of sequences in a very short period of time while remaining relatively accurate (10, 11). After the alignment is completed, three additional analysis are performed.

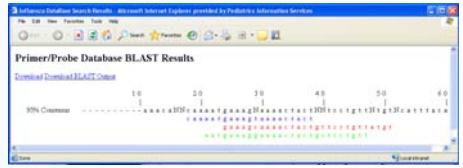
Sequence Conservation Analysis

One or more consensus sequences are calculated based on how many consensus percentages the user entered. For each consensus sequence a diagram of conserved regions is also determined. The diagram of conserved regions is designed to eliminate small unconserved areas in the consensus sequence so it is easier for the user to manually find primers and probes. Also included in the output for each position is the most conserved nucleotide, a bar graph of the percent conservation, the numerical value of the percent conservation, the percent conservation of each nucleotide, a bar graph of the gap fraction, and the numerical value of the gap fraction. Additionally, a file containing the consensus and conserved sequences in FASTA format and the alignment file can be downloaded.



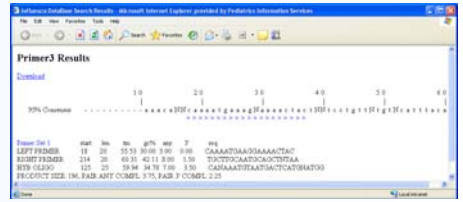
Primer/Probe Database BLAST Analysis

An additional database of influenza primers and probes was generated by manually searching the literature, extracting information from the VirOligo website, and inputting primers and probes used in our laboratory. After the consensus sequences are determined they are BLASTed against the database to help identify primers and probes, which already have experimental data that may be used with the consensus sequence. These results are displayed to the user with each hit aligned with the consensus sequence and linked to more information about the primer or probe.



Primer3 Analysis

Each consensus sequence is also sent through a Primer3 analysis. Primer3 is a program that is commonly used for primer and probe design. This program allows the user to set multiple parameters like product size, melting temperature, and GC content. When the analysis is complete the program's output is displayed to the user.



CONCLUSIONS

While there are several other websites currently available that provide access to the vast amounts of influenza sequence data, none of them provide the right tools to properly facilitate primer and probe design. Through our website we hope to provide a combination of analyses that will significantly decrease the time and money that researchers must utilize in order to develop highly effective primers and probes for influenza diagnostics. The website (irwin.hmgc.MCW.edu/IPDR) is able to align and analyze thousands of sequences in a matter of minutes (up to the number of sequences in the entire database) by utilizing a fast alignment algorithm (MAFFT). The results of the analyses are then displayed in a easy to read format and result files are provided in common formats so that they can easily be imported into other programs if the user desires. This includes a single source for all publicly available primer and probe designs to facilitate broad dissemination of knowledge. We believe this website will advance the development of useful diagnostic assays for influenza, decrease expenditures around the world, and allow for rapid response to newly emerging influenza strains or during the next Pandemic.

REFERENCES

1. **Ellis, J.S., and Zambon, M.C.** 2002. Molecular diagnosis of influenza. Rev. Med. Virol. 12:375-389.
2. **Hoffmann, E., Stech, J., Guan, Y., Webster, R.G., and Perez, D.R.** 2001. Universal primer set for the full length amplification of all influenza A viruses. Arch. Virol. 146:1-15.
3. **Templeton, K.E., Scheltinga, S.A., Beersma, M.F.C., Kroes, A.C.M., and Claas, E.C.J.** 2004. Rapid and sensitive method using multiplex real-time PCR for diagnosis of infections by influenza A and influenza B viruses, respiratory syncytial virus, and parainfluenza viruses 1, 2, 3, and 4. J. Clin. Microbiol. 42:1564-1569.
4. **Zou, S.** 1997. A practical approach to genetic screening for influenza virus variants. J. Clin. Microbiol. 35:2623-2627.
5. **Mehlmann, M., Dawson, E.D., Townsend, M.B., Smagala, J.A., Moore, C.L., Smith, C.B., Cox, N.J., Kuchta, R.D., and Rowlen, K.L.** 2006. Robust sequence selection method used to develop the FluChip diagnostic microarray for influenza. J. Clin. Microbiol. 44:2857-62.
6. **Katoh, K., Kuma, K., Toh, H., and Miyata, T.** 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511-518.
7. **Katoh, K., Misawa, K., Kuma, K., and Miyata, T.** 2002. MAFFT: a novel method to rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059-3066.
8. **Altschul, S. F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.
9. **Rozen, S. and Skaletsky, H.J.** 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S., Misener, S. (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp 365-386.
10. **Nuin, P.A., Wang, Z., and Tiller, E.R.** 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 7:471.
11. **Ahola, V., Aittokallio, T., Vihinen, M., and Uusipaikka, E.** 2006. A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics. 7:484.