Large-Scale Nucleotide Sequence Alignment for Universal Screening PCR Assay Design

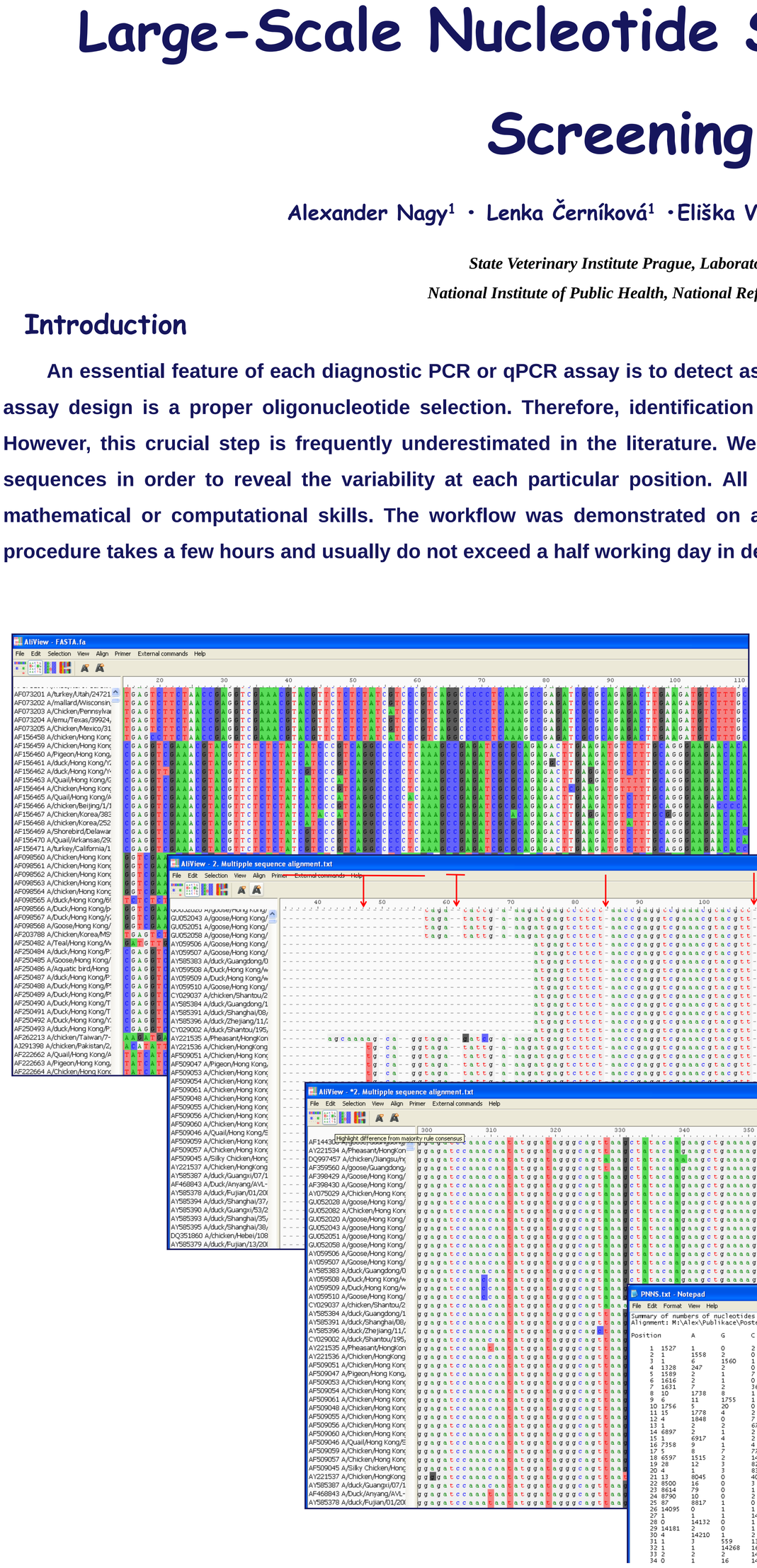
Alexander Nagy[1] · Lenka Černíková[1] ·Eliška Vitásková[1] · Tomáš Jiřinec[2] · Helena Jiřincová[2] · Martina Havlíčková[2]

State Veterinary Institute Prague, Laboratory of Molecular Methods, Sídlištní 136/24, 165 03 Prague 6, Czech Republic[1]

National Institute of Public Health, National Reference Laboratory for Influenza, Šrobárová 48, 100 42 Prague 10, Czech Republic[2]

Introduction

An essential feature of each diagnostic PCR or qPCR assay is to detect as many genetic variants of a given virus or bacterial taxon as possible. A crucial step in such a screening assay design is a proper oligonucleotide selection. Therefore, identification of the evolutionarily most conserved regions of the targeted virus genome is of utmost importance. However, this crucial step is frequently underestimated in the literature. We set up a simple workflow how to align, process, and evaluate a huge set of homologous nucleotide sequences in order to reveal the variability at each particular position. All of the steps were performed by using an ordinary desktop computer without the need for extensive mathematical or computational skills. The workflow was demonstrated on aligning more than 15.000 avian influenza A virus (IAV) M segment nucleotide sequences. The whole procedure takes a few hours and usually do not exceed a half working day in dependence on the internet connection speed and server occupancy respectively.



1. Sequence download

- The entire avian IAV M segment nucleotide sequence content of the Influenza Virus Sequence Database [1] was downloaded. The data contained 15.622 sequences (14-Jul-2015) with the corresponding FASTA file of 15,5 Mb. The unaligned sequences were shown by using the AliView program [2].

2. Multiple sequence alignment (MSA)

- For MSA the online version of the MAFFT v. 7 (Multiple Alignment using Fast Fourier Transform [3]. was implemented. The sequences were compared relative to the reference sequence.
- The figure illustrates a row alignment by using the AliView software [2].

3. Alignment validation

- Preparation of the final alignment requires alignment processing: (editing, trimming, removal of non informative, short, and low quality sequences, and restoration of the alignment continuity (Fig. 2 red arrows). This step was performed in AliView software [2].

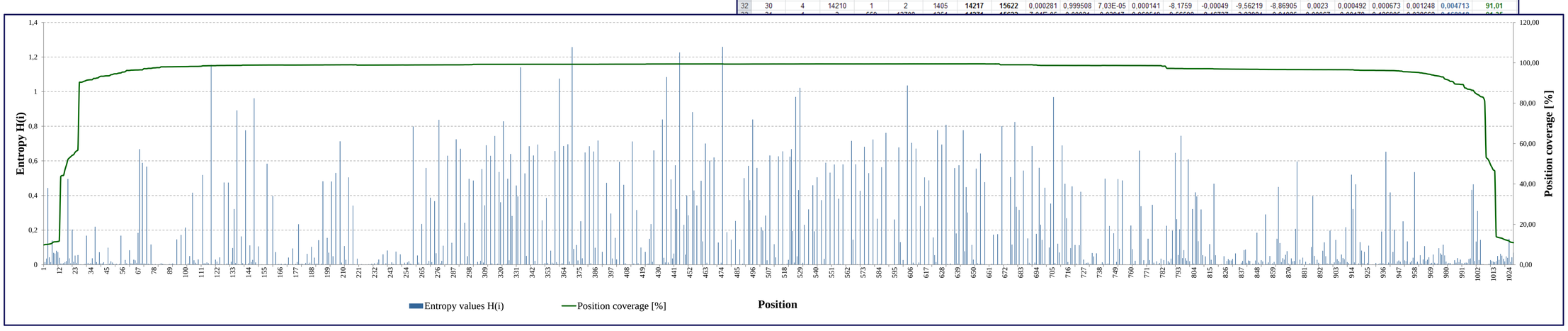
4. Positional nucleotide numerical summary (PNNS) calculation

- PNNS - counting the exact number of informative positions (i.e., A, T(U), C, G and gaps) in each column of the alignment by using the BioEdit v.7 program [4].

5. Estimation of sequence variability

- Next, the PNNS was converted to entropy values. The entropy quantifies the amount of variability in each column of the alignment. In this step the EntropyCalculator software [5] was implemented.
- The figure shows the EntropyCalculator output which includes the original PNNS (columns A-F), sum of the informative positions (column G), the entropy calculation chain (columns I-T) as well as the entropy and % coverage values (columns U and V).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Position	A	G	C	T	GAP	SUM	Total	Afreq	Gfreq	Cfreq	Ufreq	InAfreq	InGfreq	InCfreq	InUfreq	H(A)	H(G)	H(C)	H(U)	H(i)	%Coverage
2																						
3	1	1527	1	0	2	14092	1530	15622	0,998039	0,000654	0	0,001307	-0,00196	-7,33302	0	-6,63988	0,001959	0,004793	0	0,00868	0,015431	9,79
4	2	1	1558	2	0	14061	1561	15622	0,000641	0,998078	0,001281	0	-7,35308	-0,00192	-6,65993	0	0,00471	0,00192	0,008533	0	0,015163	9,99
5	3	1	6	1560	1	14054	1568	15622	0,000638	0,003827	0,994898	0,000638	-7,35756	-5,5658	-0,00512	-7,35756	0,004692	0,021298	0,005089	0,004692	0,035771	10,04
6	4	1328	247	2	0	14044	1577	15621	0,842105	0,156627	0,001268	0	-0,17185	-1,85389	-6,67013	0	0,144716	0,290369	0,008459	0	0,443544	10,10
7	5	1589	2	1	7	14023	1599	15622	0,993746	0,001251	0,000625	0,004378	-0,00627	-6,68399	-7,37713	-5,43122	0,006234	0,00836	0,004614	0,023776	0,042985	10,24
8	6	1616	2	1	0	14003	1619	15622	0,998147	0,001235	0,000618	0	-0,00185	-6,69642	-7,38956	0	0,001851	0,008272	0,004564	0	0,014688	10,36
9	7	1631	7	2	36	13946	1676	15622	0,97315	0,004177	0,001193	0,02148	-0,02722	-5,47826	-6,73102	-3,84065	0,026486	0,022881	0,008032	0,082496	0,139895	10,73
10	8	10	1738	8	1	13865	1757	15622	0,005692	0,989186	0,004553	0,000569	-5,16878	-0,01087	-5,39192	-7,47136	0,029418	0,010755	0,024551	0,004252	0,068976	11,25
11	9	6	11	1755	1	13849	1773	15622	0,003384	0,006204	0,989848	0,000564	-5,68867	-5,08253	-0,0102	-7,48043	0,019251	0,031533	0,010101	0,004219	0,065104	11,35
12	10	1756	5	20	0	13841	1781	15622	0,985963	0,002807	0,01123	0	-0,01414	-5,87549	-4,4892	0	0,013938	0,016495	0,050412	0	0,080645	11,40
13	11	15	1778	4	2	13823	1799	15622	0,008338	0,988327	0,002223	0,001112	-4,78694	-0,01174	-5,10869	-6,80184	0,039913	0,011605	0,013582	0,007562	0,072662	11,52
14	12	4	1848	0	7	13763	1859	15622	0,002152	0,994083	0	0,003765	-6,1415	-0,00593	0	-5,58188	0,013215	0,0059	0	0,021018	0,040133	11,90
15	13	1	2	2	6798	8819	6803	15622	0,000147	0,000294	0,000294	0,999265	-8,82512	-8,13197	-8,13197	-0,00074	0,001297	0,002391	0,002391	0,000735	0,006813	43,55
16	14	6897	2	1	2	8720	6902	15622	0,999276	0,00029	0,000145	0,00029	-0,00072	-8,14642	-8,83957	-8,14642	0,000724	0,002361	0,001281	0,002361	0,006726	44,18
17	15	1	6917	4	2	8698	6924	15622	0,000144	0,998989	0,000578	0,000289	-8,84275	-0,00101	-7,45645	-8,1496	0,001277	0,00101	0,004308	0,002354	0,008949	44,32
18	16	7358	9	1	4	8250	7372	15622	0,998101	0,001221	0,000136	0,000543	-0,0019	-6,70822	-8,90544	-7,51915	0,001897	0,00819	0,001208	0,00408	0,015375	47,19
19	17	5	8	7	7703	7899	7723	15622	0,000647	0,001036	0,000906	0,99741	-7,34252	-6,87252	-7,00605	-0,00259	0,004754	0,007119	0,00635	0,002586	0,020809	49,44
20	18	6597	1515	2	14	7492	8128	15620	0,811639	0,186393	0,000246	0,001722	-0,2087	-1,6799	-8,30992	-6,36401	0,169389	0,313121	0,002045	0,010962	0,495516	52,04
21	19	28	12	3	8216	7363	8259	15622	0,00339	0,001453	0,000363	0,994794	-5,68685	-6,53415	-7,92045	-0,00522	0,01928	0,009494	0,002877	0,005193	0,036844	52,87
22	20	4	1	3	8352	7262	8360	15622	0,000478	0,00012	0,000359	0,999043	-7,64492	-9,03121	-7,9326	-0,00096	0,003658	0,00108	0,002847	0,000956	0,008541	53,51
23	21	13	8045	0	405	7159	8463	15622	0,001536	0,950609	0	0,047855	-6,47851	-0,05065	0	-3,03957	0,009952	0,048151	0	0,14546	0,203563	54,17
24	22	8500	16	0	3	7103	8519	15622	0,99777	0,001878	0	0,000352	-0,00223	-6,27747	0	-7,95144	0,002228	0,01179	0	0,0028	0,016818	54,53
25	23	8614	79	0	1	6928	8694	15622	0,990798	0,009087	0	0,000115	-0,00924	-4,70094	0	-4,70094	0,09159	0,042716	0	0,001043	0,052919	55,65
26	24	8790	10	0	2	6820	8802	15622	0,998637	0,001136	0	0,000227	-0,00136	-6,78015	0	-8,38959	0,001362	0,007703	0	0,001906	0,010972	56,34
27	25	87	8817	1	0	6717	8905	15622	0,00977	0,990118	0,000112	0	-4,62846	-0,00993	-9,09437	0	0,045219	0,009833	0,001021	0	0,056073	57,00
28	26	14095	0	1	1	1525	14097	15622	0,999858	0	7,09E-05	7,09E-05	-0,00014	0	-9,55372	-9,55372	0,000142	0	0,000678	0,000678	0,001497	90,24
29	27	1	1	1	14109	1510	14112	15622	7,09E-05	7,09E-05	7,09E-05	0,999787	-9,55478	-9,55478	-9,55478	-0,00021	0,000677	0,000677	0,000677	0,000213	0,002244	90,33
30	28	0	14132	0	1	1489	14133	15622	0	0,999929	0	7,08E-05	0	-7,1E-05	0	-9,55627	0	7,08E-05	0	0,000676	0,000747	90,47
31	29	14181	2	0	1	1438	14184	15622	0,999788	0,000141	0	7,05E-05	-0,00021	-8,86672	0	-9,55987	0,000211	0,00125	0	0,000674	0,002136	90,80
32	30	4	14210	1	2	1405	14217	15622	0,000281	0,999508	7,03E-05	0,000141	-8,1759	-0,00049	-9,56219	-8,86905	0,0023	0,000492	0,000673	0,001248	0,004713	91,01



6. Results

- Finally, the entropy and %coverage values (Fig. 5, columns U and V) were summarised in a chart where the entropy was drawn as column diagram (blue) related to the main y axis while the % coverage was shown as line graph (green) related to the secondary y axis.
- The graph clearly shows the presence and distribution of the most variable (the highest columns) and also the conserved positions (Hi converging to zero) suitable for oligonucleotide selection.

References

1. **Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L,Tatusova T, Ostell J, Lipman D (2008).** J Virol 82:596–601. http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html
2. **Larsson A (2014).** Bioinformatics 15:3276–32780. http://www.ormbunkar.se/aliview/#DOWNLOAD
3. **Katoh K, Misawa K, Kuma K, Miyata T (2002).** Nucleic Acids Res 30:3059–3066. http://mafft.cbrc.jp/alignment/server/
4. **Hall TA (1999).** Nucl Acids Symp Ser 41:95–98. http://www.mbio.ncsu.edu/ RNaseP/info/programs/BIOEDIT/bioedit.html
5. **Nagy A, Jiřinec T, Černíková L, Jiřincová H, Havlíčková M (2015).** PCR Primer Design, Methods in Molecular Biology 1275. DOI 10.1007/978-1-4939-2365-6_4. http://www.szu.cz/entropycalculator (oneline version soon)