PrimerHunter: A Primer Selection Tool for Avian Influenza Subtype Identification*

Jorge Duitama[1], Mazhar Khan[2], Ion Măndoiu[1], Dipu Nair[2], and Craig Nelson[3]

[1] Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA
E-mail: {jduitama,ion}@engr.uconn.edu
[2] Department of Pathobiology and Veterinary Science, University of Connecticut, 61 North Eagleville Road, Unit 3089, Storrs, CT 06269-3089, USA
E-mail: {dipu.kumar,mazhar.khan}@uconn.edu
[3] Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Unit 3125, Storrs, CT 06269-3125, USA
E-mail: craig.nelson@uconn.edu

1 Introduction

Avian influenza belongs to the influenza type A genus of the Orthomyxoviridae family of RNA viruses. It is a highly mutable virus, with the Haemagglutinin (HA) and the Neuraminidase (NA) genes being the most variable. To date, 16 HA and 9 NA subtypes have been identified. Due to recent transmissions to humans and the ensuing risk for a global pandemic caused by highly pathogenic subtypes such as H5N1, there has been much recent work on developing rapid methods for detection and subtype identification of avian influenza infections. Nucleic acid based analyses such as the Polymerase Chain Reaction (PCR) are becoming the method of choice, largely replacing the much more labor and time consuming serotyping techniques.

Common primer design packages such as Primer3 [6] are not well suited for designing PCR primers for subtype identification since they seek to amplify a *known* target sequence, not an *unknown* target from the set of highly variable sequences that comprise a subtype. The high sequence variability within subtypes also renders unfeasible "common substring" approaches to primer selection such as [2, 5]. Methods for degenerate primer selection such as those in [3, 8] can be used to ensure amplification of all (or a large fraction) of known sequences of a given subtype, but unfortunately these methods ignore primer specificity, i.e., preventing amplification of closely related virus subtypes.

In this poster we present a new tool, called PrimerHunter, that can be used for selecting highly sensitive and specific primers for avian influenza subtyping. As in [2, 5], our tool takes as input sets of both *target* and *non-target* sequences. However, instead of searching for substrings shared by the target sequences as in [2, 5, 8], or for highly conserved regions in a multiple alignment of the target sequences as in [3], PrimerHunter ensures that selected primers amplify all

* Work supported in part by NSF awards IIS-0546457 and DBI-0543365.

target sequences and none of the non-target sequences by relying on accurate melting temperature computations based on the nearest-neighbor model of [7] and the fractional programming algorithm of [4]. Results on HA avian influenza sequences from the NCBI flu database [1] show that PrimerHunter has a high design success rate, being able to identify tens of specific primer pairs for each subtype represented in the database. Preliminary validation experiments confirm that selected primers work well in practice under the computationally predicted range of PCR conditions.

2 Problem Formulation

For simplicity, we only formalize in this section the selection of forward PCR primers; the symmetric selection of reverse primers and the formation of primer pairs are discussed in next section. We assume that all sequences are over the DNA alphabet, $\{A, C, G, T\}$, and are given in $5' - 3'$ orientation unless stated otherwise. For a sequence s, we denote by $|s|$ its length, and by $s(i, l)$ the subsequence of length l ending at position i (i.e., $s(i, l) = s_{i-l+1} \dots s_{i-1} s_i$). In our formulation we ensure that specific bases at the $3'$ end of a candidate primer match perfectly each given target sequence. This is achieved by using a 0-1 mask; for example, a mask of 1111 is used to ensure that the four bases at the $3'$ end of the primer are all perfect matches. Formally, given two $5' - 3'$ sequences of equal length p and s and a 0-1 mask M, we say that p *matches* s *according to* M if $p_i = s_i$ for every $i \in \{1, \dots, |s|\}$ for which $M_i = 1$. Here, we assume that the length of mask M is at most $|p|$ and that M is padded with 0's at the left end when it is strictly shorter than p. Given a candidate primer sequence p and a target sequence t, we denote by $\mathcal{I}(p, t, M)$ the set of positions i for which p matches $t(i, |p|)$ according to M.

The melting temperature between a $5' - 3'$ candidate primer p and a $3' - 5'$ sequence s, denoted by $T(p, s)$, is defined as the temperature at which 50% of the possible p-s duplexes are in hybridized state. We estimate $T(p, s)$ based on the nearest-neighbor model of [7] using the fractional programming algorithm of [4]. Unlike most commonly used melting temperature models, the algorithm of [4] allows accurate estimation of $T(p, s)$ for *non-complementary* sequences p and s by finding the most stable local thermodynamic alignment between them. In order to ensure sensitive PCR amplification of target sequences, we require for each selected primer p to hybridize to the $3'-5'$ complement $\bar{t}$ of each target t (a) with a melting temperature above a user specified threshold T^{min}_{target}, and (b) at a position where p matches t according to a user specified mask M. In other words, denoting by $T(p, t, M)$ the maximum over the melting temperatures $T(p, \overline{t(i, |p|)})$ for every $i \in \mathcal{I}(p, t, M)$, we require that $T(p, t, M) \geq T^{min}_{target}$ for every selected primer p and target t. Furthermore, to avoid non-specific PCR amplification, we require that each selected primer hybridizes to the $3' - 5'$ complement $\bar{t}$ of each non-target t with a melting temperature $T(p, \bar{t})$ of at most $T^{max}_{nontarget}$, where $T^{max}_{nontarget}$ is again a user specified parameter. The problem of selecting target-specific forward PCR primers is thus formulated as follows:

Given: sets $TARGETS$ and $NONTARGETS$ of target/non-target DNA sequences in $5' - 3'$ orientation, 0-1 mask M, temperature thresholds T^{min}_{target} and $T^{max}_{nontarget}$

Find: all primers p satisfying that for every $t \in TARGETS$, the set $\mathcal{I}(p, t, M)$ is not empty and the maximum melting temperature $T(p, t, M)$ is greater than or equal to T^{min}_{target} and for every $t \in NONTARGETS$ the melting temperature $T(p, \bar{t})$ is less than or equal to $T^{max}_{nontarget}$.

3 Algorithm and Implementation Details

PrimerHunter works in two stages: in the first stage forward and reverse primers are selected according to the problem formulation given in previous section, while in the second stage feasible primer pairs are formed using the primers selected in first stage.

The first stage starts with a preprocessing step that builds a hash table storing all occurrences in the target sequences of "seed" nucleotide patterns consistent with the given mask M. This is done by aligning the mask M at every position i of every target sequence t, and storing in the hash table an occurrence of the seed pattern created by extracting from $t(i, |M|)$ the nucleotides that appear at positions aligned with the 1's of M. For example, if $M = 11011$ and $t(i, 5) = GATTC$, we store in the hash table an occurrence of seed $GATC$ at position i of t.

Once the hash table is constructed, candidate primers are generated by taking substrings with lengths within a user-specified interval from one or more of the target sequences. Similar to the Primer3 package [6], PrimerHunter filters the list of primer candidates by enforcing user-specified bounds on GC Content, $3'$ GC Clamp, maximum number of consecutive mononucleotide repeats, and self complementarity. For each surviving candidate p, PrimerHunter uses the hash table to recover for each target t the list of positions $\mathcal{I}(p, t, M)$ at which p matches t according to M. It then computes the melting temperature of p with the complement of t at each of these positions, retaining p only if $T(p, t, M) \geq T^{min}_{target}$. Finally, PrimerHunter computes the maximum melting temperature between p and the complements of non-target sequences, retaining p only if $T(p, \bar{t}) \leq T^{max}_{nontarget}$ for every non-target sequence t.

The above process is repeated on the reverse complements of target and non target sequences to generate reverse primers. Then, in the second stage of the algorithm, the lists of selected forward and reverse primers are used to create feasible primer pairs by enforcing user-specified constraints on amplicon length, allowable complementarity between the two primers, and maximum difference between their melting temperatures with target sequences.

4 Experimental Results

Primer Hunter has been implemented in C++ on a standard Linux platform. We designed primer pairs for H1 to H16 subtypes using the complete different

Avian influenza HA sequences from North America available in the NCBI flu database [1] as of March 2008 (a total of 574 HA sequences). When designing primers for each subtype H_i we used all available sequences for H_i as targets, and all NCBI HA sequences with different subtypes as non-targets. Primer selection was performed using the following parameters: primer length between 20 and 25, amplicon length between 75 and 200, GC content between 25% and 75%, maximum mononucleotide repeat of 5, $3'$-end match mask of 11, no required $3'$ GC Clamp, primer concentration of $0.8\mu M$, salt concentration of $50mM$, and $T^{min}_{target} = T^{max}_{nontarget} = 40°C$. The numbers of identified primer pairs using these parameters are summarized in Table 1. Preliminary PCR experiments with three H7-specific primer pairs yielded robust amplification from H7 isolates and no amplification above the background level from H3 and H5 isolates.

Subtype	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H16
Targets	48	41	72	67	69	100	55	9	23	16	45	15	10	4
Forward Primers	52	40	39	253	61	39	67	492	143	250	263	468	43	378
Reverse Primers	53	42	67	209	64	25	79	468	145	299	247	485	33	359
Pairs	19	91	86	1717	53	3	173	10076	923	2869	2961	10908	48	6644

Table 1. Primers found for each subtype of Avian influenza HA

References

1. Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. The influenza virus resource at the national center for biotechnology information. *J. Virol.*, 82(2):596–601, 2008.
2. S. Emrich, M. Lowe, and A. Delcher. Probemer: a web-based software tool for selecting optimal DNA oligos. *Nucleic Acids Research*, 31(13):3746–3750, 2003.
3. O. Jabado, G. Palacios, V. Kapoor, J. Hui, N. Renwick, J. Zhai, T. Briese, and W.I. Lipkin. Greene scprimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Research*, 34(22):6605–6611, 2006.
4. M. Leber, L. Kaderali, A. Schonhuth, and R. Schrader. A fractional programming approach to efficient DNA melting temperature calculation. *Bioinformatics*, 21(10):2375–2382, 2005.
5. A. Phillippy, J. Mason, K. Ayanbule, D. Sommer, E. Taviani, A. Huq, R. Colwell, I. Knight, and S. Salzberg. Comprehensive DNA signature discovery and validation. *PLOS Computational Biology*, 3(5):0887–0894, 2007.
6. S. Rozen and H.J. Skaletsky. Primer3 on the WWW for general users and for biologist programmers. In S. Krawetz and S. Misener, editors, *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pages 365–386. Humana Press, Totowa, NJ, 2000.
7. J. SantaLucia and D. Hicks. The thermodynamics of DNA strucutural motifs. *Annual Review of Biophysics and Biomolecular Structure*, 33:415–440, 2004.
8. X. Wei, D. Khun, and G. Narasimhan. Degenerate primer design via clustering. In *Proceedings of the IEEE Computer Society Bioinformatics Conference*, pages 75–83, 2003.