Inference of Geographic Transmission Probability of Influenza Viruses from a Large Phylogenetic Tree

Fumiaki YANAGIHASHI[1] **Kimihito ITO**[2] **Hiroki ARIMURA**[1]

{f-yanagihashi,arim}@ist.hokudai.ac.jp itok@czc.hokudai.ac.jp

[1] Graduate School of IST, Hokkaido Univ., N14 W9, Sapporo, Japan

[2] Research Center for Zoonosis Control, Hokkaido Univ., N20 W10, Sapporo, Japan

Keywords: dynamic programming, optimal tree label assignment, Bayesian inference

1 Introduction

In this paper, we infer the Geographic Transmission Probability of influenza viruses from a large phylogenetic tree. For this purpose, firstly, we define a statistical model for generating geographic label assignments for a given phylogenetic tree. Then, we address the optimal tree label assignment problem (OTLAP) using the MAP-estimation. Secondly, we present an efficient dynamic programming algorithm DPAO that solves the OTLAP in $O(km^2n)$ time for an input tree with maximum degree k consisting of n nodes, and a $m \times m$ cost matrix over a label alphabet of size m. Finally, we apply our algorithm to the OTLAP for the phylogenetic tree of influenza viruses, then we estimate the Geographic Transmission Probability of influenza viruses.

2 Modeling and Algorithm

We define a generative model for geographic label assignments to the nodes of a given phylogenetic tree (Figure 1) . Let $\Sigma = \{1, \ldots, m\}$ be an alphabet of m geographic labels such as `Asia, North-America, South-America, Oceania` and `Europe` at which virus sequences are sampled. Suppose that $\mathcal{T}$ is a phylogenetic tree with n_1 internal nodes in $V_I = \{1, \ldots, n_1\}$ and n_2 leaves in $V_L = \{n_1+1, \ldots, n_1+n_2\}$. Let $n = n_1+n_2$ be the total size of $\mathcal{T}$. Internal and external label assignments are mappings $f : V_I \to \Sigma$ and $g : V_L \to \Sigma$ that assign geographic labels of Σ to internal nodes and leaves, respectively.

Let $\mathbf{x} = (x_i)_{i=1}^{n_1}$ and $\mathbf{z} = (z_j)_{j=n_1+1}^{n_1+n_2}$ be the vectors of n_1 *internal labels* and n_2 *external labels*, respectively. Each x_i $(i \in V_I)$ and z_j $(j \in V_L)$ indicate the labels assigned by f and g to the i-th internal node and the j-th leaf of $\mathcal{T}$. Let Θ be a set of parameters that define the probability $q(x_0 \mid \Theta)$ of the label $x_0 \in \Sigma$ of the root, and the conditional probability $p(x \mid y, \Theta)$ for the label x of node i and the label y of its parent $\pi(i)$. Then, our probabilistic model is given by the joint probability distribution for $\mathbf{x}$ and $\mathbf{z}$:

$$p(\mathbf{x}, \mathbf{z} \mid \Theta) \quad = \quad q(x_0 \mid \Theta) \prod_{i=1}^{n_1} p(x_i \mid x_{\pi(i)}, \Theta) \prod_{j=n_1+1}^{n_1+n_2} p(z_j \mid x_{\pi(j)}, \Theta) \tag{1}$$

The input of the problem is a triple $\mathcal{I} = (\mathcal{T}, g, \Theta)$. The task is the MAP-estimation (maximum a posteriori probability estimation) of internal labeling f, that is, to find a labeling $\mathbf{x}$ that maximizes the a posteriori probability $p(\mathbf{x} \mid \mathbf{z}, \Theta)$. We present an efficient algorithm DPAO (Dynamic Programming Algorithm for OTLAP) for the problem. To perform MAP-estimation, we transform Eq. (1) to the score function in additive form $S(\mathbf{x}, \mathbf{z} \mid \Theta) = -\log p(\mathbf{x}, \mathbf{z} \mid \Theta) = I_{x_0} + \sum_{i=1}^{n_1} D_{x_i, x_{\pi(i)}} + \sum_{j=n_1+1}^{n_1+n_2} D_{z_j, x_{\pi(j)}}$, where $I_x = -\log q(x \mid \Theta)$ and $D_{x,y} = -\log p(x \mid y, \Theta)$. Given $\mathbf{z}$, DPAO finds the best $\mathbf{x}$ that minimizes

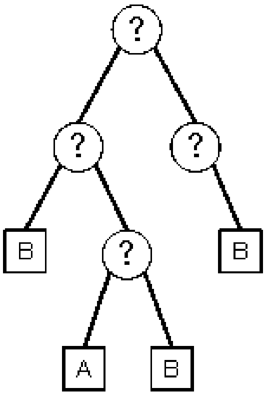

Figure 1: A phylogenetic tree with labeled leaves

Table 1: The Estimated Geographic Transmission Probability. Row variables represent parent node's geographic label and the column variables represent child node's geographic label. EA, E, NA, O, CA, SA, A, and ME are abbreviations for E-SE-Asia, Europe, N-America, Oceania, C-Asia, S-America, Africa and Middle-East respectively.

From \ To	EA	E	NA	O	CA	SA	A	ME
EA	0.878	0.027	0.048	0.027	0.005	0.01	0.004	0.001
E	0.045	0.842	0.062	0.025	0.005	0.017	0.004	0.000
NA	0.029	0.042	0.882	0.018	0.001	0.016	0.006	0.006
O	0.041	0.041	0.04	0.864	0.001	0.012	0.000	0.001
CA	0.026	0.053	0.000	0.026	0.869	0.000	0.026	0.000
SA	0.036	0.044	0.085	0.022	0.000	0.81	0.003	0.000
A	0.05	0.125	0.1	0.000	0.000	0.05	0.675	0.000
ME	0.125	0.000	0.125	0.000	0.125	0.000	0.000	0.625

the score $S(\mathbf{x}, \mathbf{z} \,|\, \Theta)$ by dynamic programming with table DP Table $BS[v][b]$ for all combinations of node v and label b, which stores the minimum score of the subtree $\mathcal{T}(v)$ with the root label $b \in \Sigma$ over all assignments. The subprocedure ComputeTable computes DP Table from leaf to root as follows. If v is a leaf, $BS[v][b]$ is 0 if $g(v) = b$ and ∞ otherwise. If v is internal node, ComputeTable firstly call itself recursively to compute the DP Table for all of v's children. Then, it computes $BS[v][a] = \sum_{i=1}^{k} BS_i^a$, where v_i is the i-th child of v and $BS_i^a = min\{D_{ab} + BS[v_i][b] \mid b \in \Sigma\}$. Finally, subprocedure TraceBack computes from root to leaf the optimal tree label assignment $\mathbf{x}$ with DP Table. It is not hard to see that DPAO computes OTLAP in polynomial time for node.

3 Experimental Result and Discussion

A total of 3791 HA influenza virus sequences were downloaded from NCBI Influenza Virus Resource. Then, a phylogenetic tree was constructed by neighbor-joining method [1]. Each leaf node has a geographic label indicating the area where the virus was isolated. The geographic labels of internal nodes are inferred by DPAO algorithm. For each pair of areas, the number of edges that have the areas at their ends are counted. The Geographic Transmission Probability from area A to area B was obtained by dividing the number of those edges connecting A to B by the total number of edges connecting area A to any areas. Table 1 shows the resulting table. The asymmetricity seen in the table suggests that there are difference between geographic distribution of where viruses come from and that of where viruses go to. See [4] for the details.

References

[1] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[2] M. R. Garey, D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., 1979.

[3] C. Russell, T. Jones, I. Barr, N. Cox, R. Garten, V. Gregory, I. Gust, et al. Science, *The Global Circulation of Seasonal Influenza A (H3N2) Viruses.*, 320(5874):340–341, 2008.

[4] F. Yanagihashi, K. Ito, H.Arimura, *Optimal Label Assignment Problem for Rooted Trees and Its Application to Phylogenetic Trees*,Bioinformatics and Genomics , 2009.