Spatiotemporal Modeling to Measure the Effects of Mutations and Selection Pressures

Hirohisa Kishino[1,4], Teruaki Watabe[2], Reiichiro Nakamichi[3], and Shuichi Kitada[3]
[1] The University of Tokyo, Tokyo 113-8657 JAPAN
[2] Kochi University, Kochi 783-8505 JAPAN
[3] Tokyo University of Marine Science and Technology, Tokyo 108-0075 JAPAN
[4] Corresponding author: Hirohisa Kishino, e-mail: kishino@lbm.ab.a.u-tokyo.ac.jp

Abstract

The successive innovations of measurement technology have provided direct evidence that can be used to identify the molecular mechanisms behind biological phenomena. The findings and accompanying data are usually deposited in public databases. In this article, we report some attempts to measure adaptive mutations and selection pressures by integrating multiple sources of information through spatiotemporal statistical modeling of the likelihood and the prior distribution. In the first attempt, we quantified the chance of adaptive mutations in a viral population by integrating a population dynamics model, a formula in population genetics, biological sequence data, and information from the protein structural database. In the second attempt, we measured the response to stresses by the graphical modeling of gene expression and phenotypes. The chronological order data and the information from biological databases determine the constraints on the graph space. Instead of attempting to estimate the whole interaction network, we reconstructed a maximal connected subgraph that included either the target phenotypes or the core pathway using the maximum entropy principle.

Keywords: adaptive mutation, directed core-graph of gene expression, protein tertiary structure, selection pressure, statistical genetic modeling

1. Introduction

The biotechnology and medical sciences play significant roles in improving health and quality of life in modern societies. However, ecosystems are experiencing unprecedentedly high direct and indirect selection pressures as the result of human activities. As a consequence, ecosystems are responding to these stresses at an unprecedented pace. Sometimes, such responses cause major ecosystem changes that can become a threat to society. Therefore, it is important to detect and monitor such responses and to understand the mechanisms that drive adaptive evolution.

RNA viruses provide an ideal opportunity to quantify the effects of mutations on adaptive evolution, because they have a simple life-cycle and highly mutable genomes. The genomes of viruses contain protein coding genes but lack the transcription factors and cis-elements required for gene regulation; although microRNAs that silence target genes have been identified recently in viral genomes. When a viral genome is integrated into the genome of its host, the virus uses the host cell machinery to produce viral proteins. It is possible that structural changes in the viral proteins may be a major driving force of adaptive evolution. Structural changes in the viral proteins might affect their interactions with the other proteins or with ligands at important stages of their life-cycle.

Biological conservation and stock enhancement can change the fitness of a target species. For example, Araki *et al*. (2007) observed a decline in reproductive success among captive-bred salmon compared with among wild salmon. The reproductive success of steelhead trout was reported to vary depending, for the most part, on cross, year of release, and environmental conditions (Kitada *et al*. (2011)). Nevertheless, it is essential to understand the mechanisms behind changes in reproductive behaviors. The

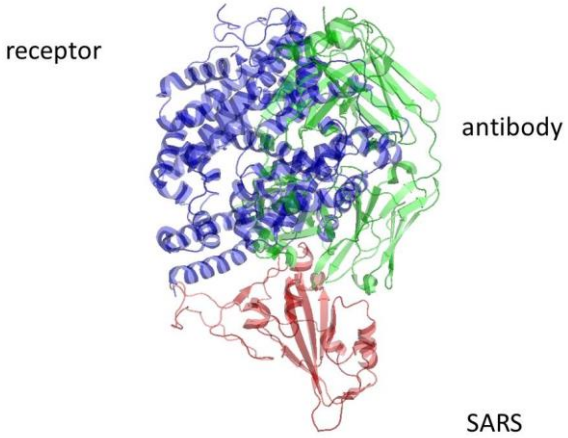

Figure 1 Antibody (green) and receptor (blue) bound to the spike protein of the SARS virus (red).

genomes of higher organisms contain coding sequences and regulatory elements. While structural changes in proteins can have a strong and mostly harmful impact on their activity, changes in gene expression can have mild effects that help generate the phenotypic diversity of a population. Because of the interactions among genes, changes in the expression level of a single gene can affect many components of the physiological and endocrine systems of an organism.

In this paper, we predict the fate of viral mutations by integrating a population dynamics model, information from the protein structure database, and population genetics theory. We also detect the evolutionary footprint of a host-pathogen arms race using protein structure information to specify the prior distribution of the key parameter in a model of the evolutionary process. We use additional information from, for example, transcription factor binding sites databases, to constrain the search space of graphs that represent the direct and indirect effects of the selection pressure.

2. Risk analysis of viral adaptation via a likelihood-based binding ability

An example of an antibody and cell-receptor together binding to a viral protein is shown in Figure 1. Because the binding regions overlap, a mutation in the viral protein that reduces its ability to bind to the antibody may also reduce its ability to bind to the cell receptor. The fitness advantage of viral mutations is measured as the expected increase of the viral load. It can be simulated using a mathematical model of the population dynamics of viruses (V), antibodies (A), normal cells (C^N) and infected cells (C^I) in a host (Nowak and Bangham (1996)) as:

$$\frac{dC^N}{dt} = \lambda - dC^N - \beta C^N V$$
$$\frac{dC^I}{dt} = \beta C^N V - aC^I$$
$$\frac{dV}{dt} = KC^I - uV - \beta C^N V - qVA$$
$$\frac{dA}{dt} = rV - hA - qVA$$

The model includes two key parameters; one is the binding ability, β, of the viral protein to the receptor of normal cells, and the other is its binding ability, q, to the antibodies. We estimated the effect of mutations on the values of these parameters by

calculating the change of sequence-structure fitness (Watabe *et al*. (2007), Watabe and Kishino (2010)). The binding ability of two proteins A and B is measured as a likelihood ratio of the two protein sequences X_A and X_B given the structures of the free proteins (Y_A and Y_B) and of the proteins in the protein complex (Y_{A+B}) as:

$$K_a \propto \frac{P(X_A, X_B \mid Y_{A+B})}{P(X_A \mid Y_A)P(X_B \mid Y_B)}.$$

Because the second order approximation, the likelihood of a sequence, $X = (a_1, \ldots, a_n)$, given the structure, Y, is calculated as:

$$P(X \mid Y) \cong \prod_{i=1}^{n} P(a_i \mid Y) \prod_{i<j} \frac{P(a_i, a_j \mid Y)}{P(a_i \mid Y)P(a_j \mid Y)}.$$

Assuming that the conditional probabilities $P(a_i \mid Y)$ and $P(a_i, a_j \mid Y)$ depend only on the local structure surrounding the amino acid residues, they are estimated as the amino acid frequencies in the proteins in the Protein Data Bank (Simons *et al*. (1999)). Using the expected value of the selective advantage, it is possible to predict the risk of viral adaptation by using population genetics theory (Wright (1931), Kimura (1962)).

3. Detecting the region under diversifying selection using a hierarchical Bayes model

Because of the arms race between virus and host, viral genomes evolve under positive diversifying selection. It is possible to estimate their evolutionary history by comparing the protein sequences. DNA triplets (the codons) in the protein coding sequences of genes are translated into the amino acids that make up protein sequences. Because of codon redundancy (64 codons code for only 20 amino acids), not all nucleotide substitutions change the amino acid. Positive selection is observed as an elevated ratio, $\omega = d_N / d_S$, of the rate of nonsynonymous (d_N) to synonymous substitutions (d_S). (A nonsynonymous substitution is a one base change that produces an amino acid change in the sequence; a synonymous substitution is a one base change that does not produce an amino acid change.) Yang *et al*. (2000) incorporated heterogeneity among sites into a likelihood model to detect positive selection and found that positive selection was acting on a small proportion of viral proteins. Because amino acid residues under positive selection are generally considered to be spatially clustered, the region under positive selection could perhaps be estimated by introducing a smoothness prior (Suzuki and Gojobori (2004)).

Assuming that the molecular evolution follows a Markov process, the transition rate from codon *i* to codon *j* at the h th site is described as:

$$q_{ij}^{(h)} = \begin{cases} 0 & \text{for more than one nucleotide substitution between } i \text{ and } j \\ \pi^{(j)} & \text{for synonymous transversion} \\ \kappa\pi^{(j)} & \text{for synonymous transition} \\ \omega_h\pi^{(j)} & \text{for non-synonymous transversion} \\ \omega_h\kappa\pi^{(j)} & \text{for non-synonymous transition} \end{cases},$$

where $\pi^{(j)}$ is the equilibrium probability and κ is the transition (A ↔ G or C ↔ T) to transversion (A ↔ C/T or G ↔ C/T) rate ratio. Given a tree topology, T, the likelihood of a set of sequences, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, is a function of the transition probabilities and the equilibrium probabilities:

$$L = \prod_{h=1}^{n} f_T(\mathbf{X}_h) = \prod_{h=1}^{n} \left[\sum_{Z_{h0}} \pi_{Z_{h0}} \prod_{v_l \in V(T) \setminus v_0} \sum_{Z_{hv_l}} P_{Z_{anc(v_l)} Z_{v_l}} \left(t_{anc(v_l), v_l} \mid \omega_h \right) \right],$$

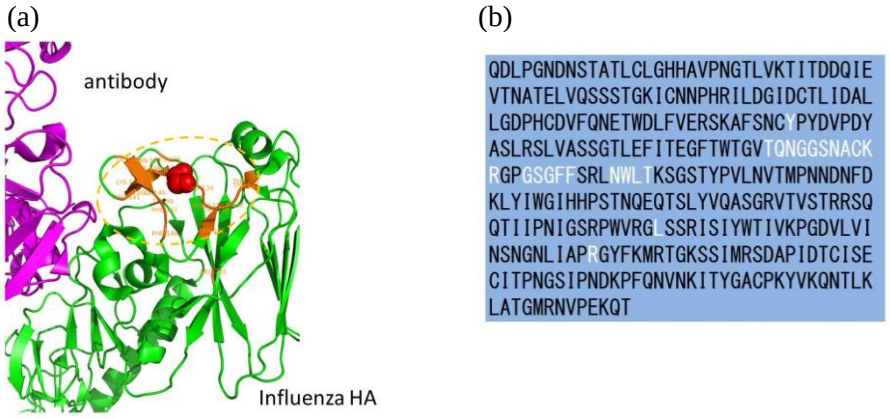

Figure 2 Antibody (pink) bound to the hemagglutinin (HA) protein of the influenza virus (green) and the HA protein sequence. The neighborhood of an amino acid in the spatial structure and in the primary sequence is shown. (a) The 10 Å neighborhood (orange) of an amino acid (red) in the spatial structure. (b) The corresponding amino acids (white) in the primary structure.

where $V(T)$ is the set of nodes of T, and $anc(v)$ represents the ancestral node of the node v. The d_N/d_S ratio has the three values of purifying selection ($\omega_1 < 1$), neutral ($\omega_2 \sim 1$), and positive selection ($\omega_3 > 1$). We introduce the Ising model as a prior distribution of the states of ratios $s_1,...,s_n$ to express the aggregated pattern of amino acid residues under positive selection as:

$$P(s_1,...,s_n) \propto \exp\left(\lambda\left(\sum_{h<h'}\left(\delta_{s_h s_{h'}} - \frac{1}{3}\right)\exp(-\alpha r_{hh'})\right)\right),$$

where $r_{hh'}$ is the spatial distance between the C^{α} atoms of the hth residue and the h'th residue. Note that, because an amino acid sequence folds to form the protein structure, neighboring amino acid residues are not clustered in the primary sequence (Figure 2). We estimate the hyperparameters λ and α by maximizing the marginal likelihood.

4. Maximal connected subgraph including the core gene expression set

Correlated gene expression is described well by graphical modeling. The graph consists of a set of nodes (or vertices), V, and a set of edges, E, that connect the nodes. The structure of the graph is represented by an adjacent matrix that specifies the presence and absence of edges between the nodes. Significant edges can be selected either by using the L_1-penalized likelihood approach (Friedman *et al*. (2008)) or by minimizing the information criteria (Edwards *et al*. (2010)). Modularity is identified as the block diagonal approximation of the adjacent matrix (Figure 3a).

Because of the high dimensionality of microarray data compared with the smaller sample size, it is important to control the signal-to-noise ratio. Instead of estimating the interaction among all the genes in a genome, we focused on a maximal connected subgraph that includes either the target phenotypes or the core pathway (Figure 3b). Additional information, such as the chronological order of temporal variation and transcription factor binding sites from public databases, is used as a constraint on the direction of the graph. The maximal connected subgraph that is obtained can quantify the biological mechanism that generates the diversity of the target phenotypes.

Assuming a directed acyclic graph as a first approximation, the likelihood of the expression profile $\{\mathbf{X}_v : v \in V\}$ given the graph structure (V,E) is:

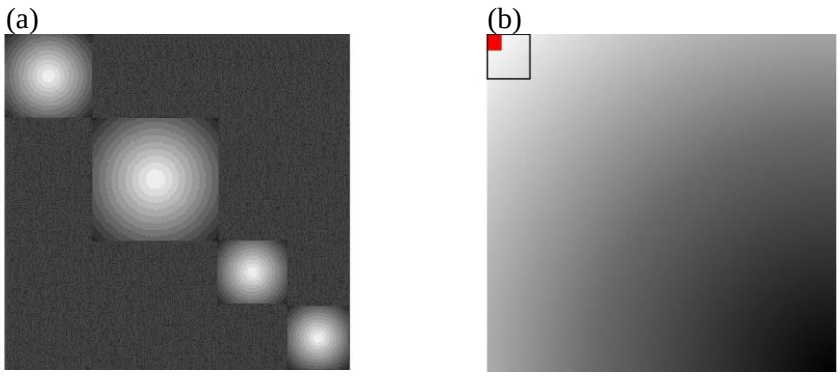

Figure 3 Schematic diagram of an adjacent matrix that describes a graph. The brightness represents the intensity of the edges. (a) Nodes are re-ordered to extract the modularity. (b) Information-based maximal connected subgraph including the core set (red).

$$L = \prod_{v \in V} P\left(\mathbf{X}_v \mid \{\mathbf{X}_{v'} : v' \in pa(v)\}\right),$$

where $pa(v)$ is the set of v's parental nodes. Starting with the core subgraph, we expand the graph as far as the additional edges that have significant mutual information. We developed a genetic algorithm with the fitness measure of AIC (Akaike (1974)) to construct the maximal connected subgraph. This subgraph consists of moving an edge (mutation), adding an edge to a terminal node (insertion), removing an edge (deletion), and crossover of edges among a parent graphs.

5. Conclusions

Modern societies are impacting on the ecosystems. The strong selection pressures that this impact produces give rise to adaptive mutations that are fixed promptly to a population. Technological innovations for measuring biological phenomena make it possible to quantify the hidden molecular mechanisms of the selection pressures and the resultant adaptive evolution. Our spatiotemporal statistical models integrate the growing knowledge of the biological sciences and the information that is contained in biological databases. These models may serve as tools to measure the effects of selection on protein sequences and structure and on the physiological systems that they control, which will help in understanding the mechanisms that guide adaptive evolution.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Araki, H., Cooper, B. and Blouin, M. S. (2007). Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild, *Science*, **318**, 100–103.

Edwards, D., de Abreu, G. C. G. and Labouriau, R. (2010) Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests, *BMC Bioinformatics*. **11**, 18.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, **9**, 432-441.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population, *Genetics*. **47**, 713–719.

Kitada, S., Kishino, H. and Hamasaki, K. (2011). Bias and significance of relative reproductive success estimates based on steelhead trout (*Oncorhynchus mykiss*) data: a Bayesian meta-analysis, *Canadian Journal of Fisheries and Aquatic Sciences*, **68**, 1827–1835.

Nowak, M. A. and Bangham, C. R. M. **(**1996**)**. Population dynamics of immune responses to persistent viruses, *Science*, **272**, 74–79.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins, *Proteins*, **34**, 82–95.

Suzuki, Y. and Gojobori, T. **(**1999**)**. A method for detecting positive selection at single amino acid sites, *Molecular Biology and Evolution*, **16**, 1315–1328.

Watabe, T. and Kishino, H. (2010). Structural considerations in the fitness landscape of a virus. *Molecular Biology and Evolution*. **27**, 1782–1791.

Watabe, T., Kishino, H., Martins, L. O. and Kitazoe, Y. (2007).A likelihood-based index of protein-protein binding affinities with application to influenza HA escape from antibodies. *Molecular Biology and Evolution*, **24**, 1627–1638.

Wright, S. (1931). Evolution in Mendelian populations, *Genetics*, **16**, 97–159.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. K. **(**2000**)**. Codon-substitution models for heterogeneous selection pressure at amino acid sites, *Genetics*, **155**, 431–449.