MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 1119–1123 doi: 10.1111/j.1755-0998.2011.03045.x

Estimating the location and shape of hybrid zones

BENJAMIN GUEDJ* and GILLES GUILLOT†
*Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, 175 rue Chevaleret 75013 Paris, France,
†Informatics and Mathematical Modelling Department, Technical University of Denmark, Richard Petersens Plads, Bygning 305, 2800 Lyngby, Copenhagen, Denmark

Abstract

We propose a new model to make use of georeferenced genetic data for inferring the location and shape of a hybrid zone. The model output includes the posterior distribution of a parameter that quantifies the width of the hybrid zone. The model proposed is implemented in the GUI and command-line versions of the Geneland program versions ≥3.3.0. Information about the program can be found on http://www2.imm.dtu.dk/~gigu/Geneland/.

Keywords: Population structure, hybridization, admixture, selection pressure, Markov chain Monte Carlo, Geneland

Received 10 February 2011; revision received 24 May 2011; accepted 6 June 2011

Background

Hybrid zones have been the object of considerable attention as they are seen as *windows on the evolutionary process* (Harrison 1990), and inference about genetic structure in their neighbourhood can provide valuable insights into the intensity of selection. This is made possible through the existence of explicit models of cline shapes as a function of selection (Haldane 1948; Bazykin 1969; Kruuk *et al.* 1999). To analyse hybrid zones, scientists have relied on a variety of approaches. They can use hybrid zone models that predict patterns of allele frequencies and fit corresponding parametric curves (Analyse program, Barton & Baird 1998) or nonparametric curves (Macholán *et al.* 2008). They can also use general-purpose computer programs such as Structure (Pritchard *et al.* 2000) that seek patterns in ancestries of individuals without reference to any model of hybrid zones. Here, we propose a new spatial model that combines features of both approaches: it explicitly accounts for the presence of a cline without making restrictive assumption about the shape of the cline path and it also retains the flexibility of the admixture model of Structure.

Model

We assume that individuals in the data set at hand have alleles with origins in K distinct gene pools characterized by different allele frequencies. We denote by $z = (z_{il})$ the matrix of genotype data where z_{il} denotes the genotype of individual i at locus l and by f_{kla} the frequency of allele a at locus l in the k-th gene pool. We introduce the matrix $q = (q_{ik})$, where q_{ik} refers to individual i's genome proportion originating from cluster k. For diploid individuals and assuming statistical independence of the two alleles harboured on the same locus of homologous chromosomes, we have

$$L(z_{il}|f,q) = \sum_{k=1}^{K} q_{ik} f_{kl z_{il1}} f_{kl z_{il2}} (2 - \delta_{z_{il1}}^{z_{il2}}) \qquad (eqn\,1)$$

where δ_a^b is the Kronecker symbol, i.e. $\delta_a^b = 1$ if $a = b$ and 0 otherwise.

For haploid data, we have

$$L(z_{il}|f,q) = \sum_{k=1}^{K} q_{ik} f_{kl z_{il}} \qquad (eqn\,2)$$

Further, assuming independence across the different loci, we have

$$L(z|f,q) = \prod_{i=1}^{n} \prod_{l=1}^{L} L(z_{il}|f,q) \qquad (eqn\,3)$$

This is the classical admixture likelihood assumed in the Structure program and related works. We assume further that each gene pool (or cluster) occupies a certain fraction of the spatial domain. The spatial domain of each cluster is assumed to display a certain organization in the sense that the various clusters do not overlap too much in space. This is accounted for by a so-called coloured Poisson–Voronoi tessellation, which is the spatial model implemented in the Geneland program. An example is given on Fig. 1. The reader unfamiliar with this model is invited to refer to Guillot *et al.* (2005) and Guillot *et al.* (2009) for a detailed presentation. See also Appendix (in

Correspondence: Gilles Guillot, E-mail: gigu@imm.dtu.dk

particular Figure 5) for details about how the novel part of the model connects to earlier versions of the Geneland program.

The model introduced here differs from earlier versions of Geneland in that it models admixture and from Structure in that it is spatial. Those two features are accounted for as follows: each vector of admixture proportions $q_{i.} = (q_{ik})_{k=1,\ldots,K}$ is assumed to follow a Dirichlet distribution $\mathscr{D}(\alpha_{i1}, \ldots, \alpha_{iK})$. We denote by d_{ik} the distance of individual i to cluster k (in particular, $d_{ik} = 0$ if individual i has been sampled in cluster k), and we assume a deterministic relationship

$$\alpha_{ik} = a \exp(-d_{ik}/b) \qquad (eqn\,4)$$

By a standard property of the Dirichlet distribution, under eqn (4), the expected value of q_{ik} is

$$E[q_{ik}] = \frac{e^{-d_{ik}/b}}{\sum_k e^{-d_{ik}/b}} \qquad (eqn\,5)$$

In the presence of $K = 2$ clusters in contact along a hybrid zone, and if individual i belongs to cluster 1, then by definition $d_{i1} = 0$, and we get

$$\begin{aligned} E[q_{i1}] &= \frac{e^{-d_{i1}/b}}{e^{-d_{i1}/b} + e^{-d_{i2}/b}} \\ &= \frac{1}{1 + e^{-d_{i2}/b}} \end{aligned} \qquad (eqn\,6)$$

i.e. the well-known sigmoid function (or logistic function, cf, e.g. Cramer 2003) familiar to people studying hybrid zones, which is also equivalent to the hyperbolic tangent cline model described by Bazykin (1969):

$$\frac{1}{2}(1 + \tanh(d)) = \frac{1}{1 + e^{-2d_{i2}/b}} \qquad (eqn\,7)$$

Under this model, the width of the cline (defined as the inverse of the maximum gradient) is $w = 4b$. The variation of the expected admixture coefficients is illustrated in Fig. 2.

Parameter a is a-dimensional, it does not affect the expected value of q_{ik} but controls its variance with $V[q_{ik}] \propto 1/a$. Large a values correspond to data sets with individuals displaying pretty similar admixture proportions within clusters. Parameter b is a spatial scale parameter, it has the dimension of a distance and is expressed in the same unit as spatial coordinates. Large b values correspond to situations where admixture coefficients are loosely structured in space. At the limit where $b = +\infty$, the vector $q_{i.}$ follows a flat Dirichlet distribution and the model does not display spatial features at all. Conversely, at the limit value $b = 0$, all individuals display admixture proportions that are 0 or 1 with a spatial pattern mirroring exactly the underlying Poisson–Voronoi tessellation. In all subsequent analyses and in our program, we place a uniform prior on a and b and assume independence of these two parameters. See Appendix for details on the inference algorithm.

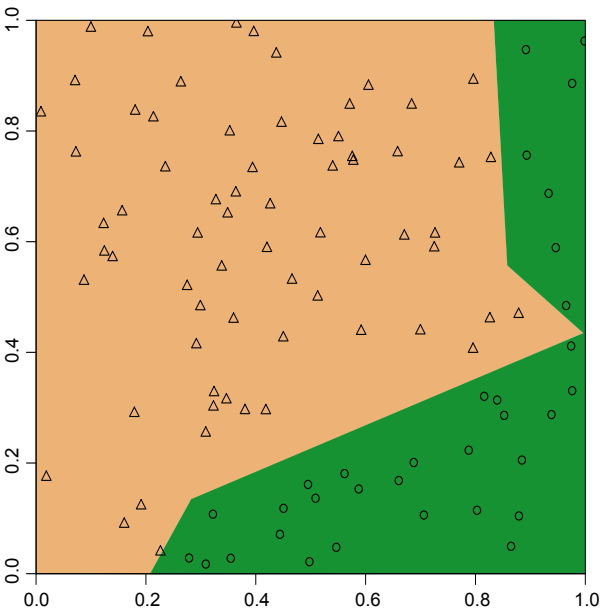

Fig. 1 Example of $K = 2$ spatial clusters simulated from a coloured Poisson–Voronoi prior model. The points represent putative sampling sites of individuals (symbols' shapes represent cluster membership). The realization of the Poisson process governing the tessellation is not shown for clarity.

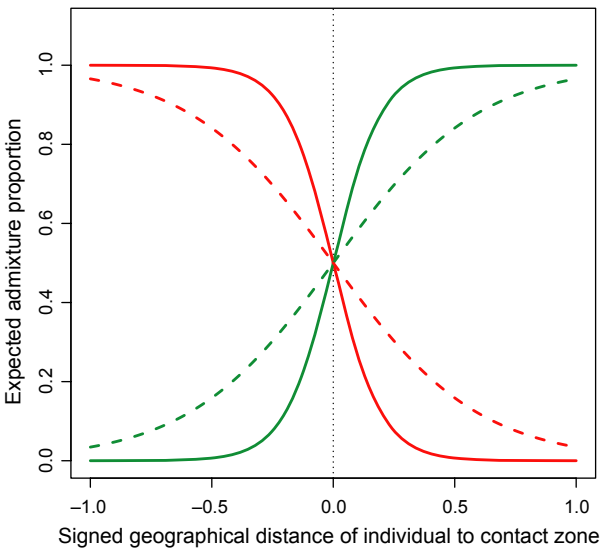

Fig. 2 Examples of spatial variation of expected admixture proportions in the presence of two clusters. Individuals whose proportions are displayed here are assumed to be continuously located along a linear transect crossing perpendicularly a hybrid zone. Decreasing curves: expected admixture proportion q_{i1}. Increasing curves: expected admixture proportion q_{i2}. Continuous lines: $a = 1$, $b = 0.1$, dashed lines: $a = 1$, $b = 0.3$. Note that the curves are exactly sigmoid (logistic) functions.

Test of the method on simulated data

To test the efficiency of our approach, we carried out MCMC inference on data produced by simulation under our model. We explored various situations in terms of variance ($\propto 1/a$) and spatial scale (b) of admixture proportions but also in terms of number of loci L. In all cases, the data set consists of 200 individuals belonging to $K = 2$ different clusters located on a $[0, 1] \times [0, 1]$ square. We explored a broad range of pairwise cluster differentiations as measured by F_{ST} (lower quartile $F_{ST} = 0.003$, upper quartile 0.03). Some graphical examples of inference are presented in Figs 3 and 4. Our main numerical results are summarized in Table 1. It appears that our method is accurate even for moderate to small numbers of loci ($L = 20$ or $L = 10$). We also note that the accuracy decreases when b increases (i.e. in case of loose spatial structure), which is the price to pay for using a spatial model. Another observed loss of accuracy (not shown here) occurs when the spatial scale of the cline is smaller than the resolution of the spatial sampling. In the extreme case when the width of the cline is smaller than the smallest interdistance between individuals, no reliable inference of b can be made. This means that users must have an idea of the characteristic scale of the cline before sampling.

Discussion

Our model of clinal variation is the same model as in MacCallum *et al.* (1998), the equivalent options in Analyse are for 2D spatial analysis with constant cline width along the course of the zone centre and a sigmoid cline cross section. The difference between this existing method and our global model is that the former is constrained by a very simple model of the path of the zone centre through space, the limitations of which are discussed at length by Bridle *et al.* (2001). This difference highlights one of the properties of our work: placing an explicit clinal admixture model in the context of the Geneland Voronoi tessellation approach—which is reminiscent of the approach taken by Macholán *et al.* (2011)—removes the existing unrealistic restriction for modelling the course of a hybrid zone centre through a 2D field area (although it does not allow for cline width to vary along the course of the hybrid zone).

Analyse requires the user to a priori reduce multi-allelic loci to two states, corresponding to origin in two source clusters. The frequency of these two states in the source clusters can be co-estimated with cline parameters; however, the reduction to two states very much reduces Analyse's applicability to, for example, microsatellite data, as a posteriori the user cannot, for example, quantify which allelic states are most associated with each source. In contrast, Structure co-assigns allelic state to source while estimating their frequencies in clusters, making microsatellites easy to use, but of course there is no spatial model. In this sense, our work combines aspects of each approach, allowing frequencies of multi-allelic allelic states to be co-estimated with cline parameters in a spatial explicit way.

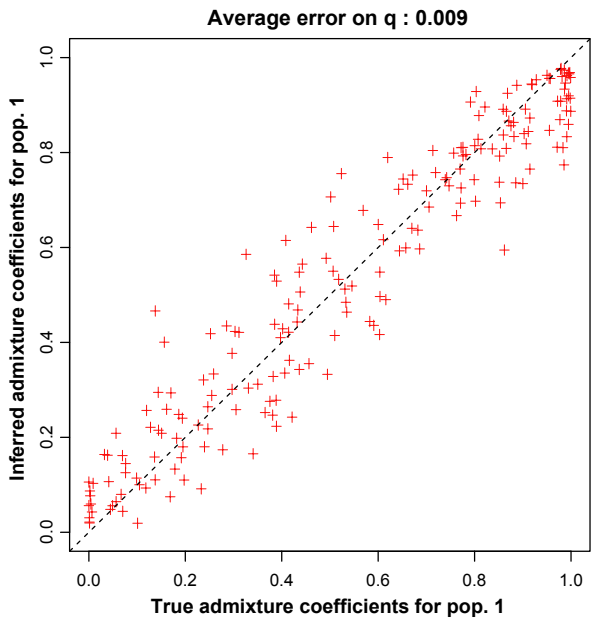

Fig. 3 Examples of result of inference: estimated versus true admixture proportions. The hyperparameters of the admixture proportions were $a = 1, b = 0.3$.

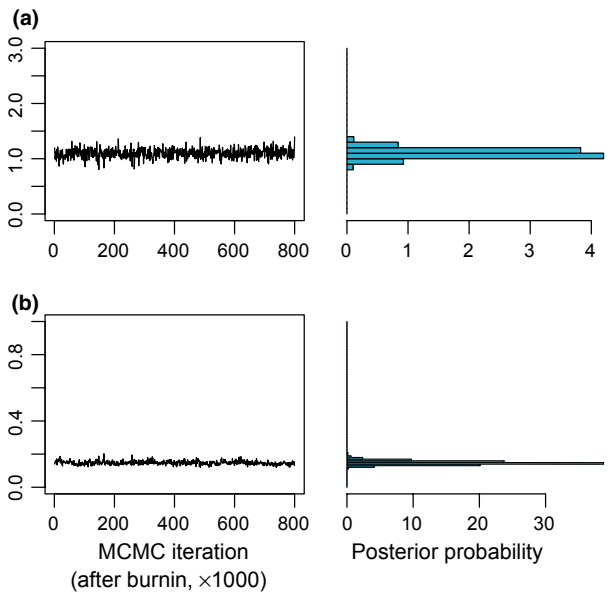

Fig. 4 Example of result of inference: MCMC trace (left) and posterior distribution (right) of parameters a and b.

b	$L = 10$			$L = 20$			$L = 50$		
	$a = 1$	$a = 2$	$a = 5$	$a = 1$	$a = 2$	$a = 5$	$a = 1$	$a = 2$	$a = 5$
0.05	0.028	0.028	0.027	0.019	0.015	0.010	0.005	0.012	0.006
0.1	0.038	0.030	0.029	0.019	0.021	0.015	0.007	0.012	0.015
0.3	0.036	0.036	0.027	0.022	0.035	0.018	0.009	0.010	0.008
0.7	0.046	0.031	0.020	0.022	0.022	0.016	0.010	0.009	0.015

Table 1 Mean square error in the inference of admixture proportions. Data are generated by simulation from our prior-likelihood model. Each number is obtained as an average over 10 independent data sets

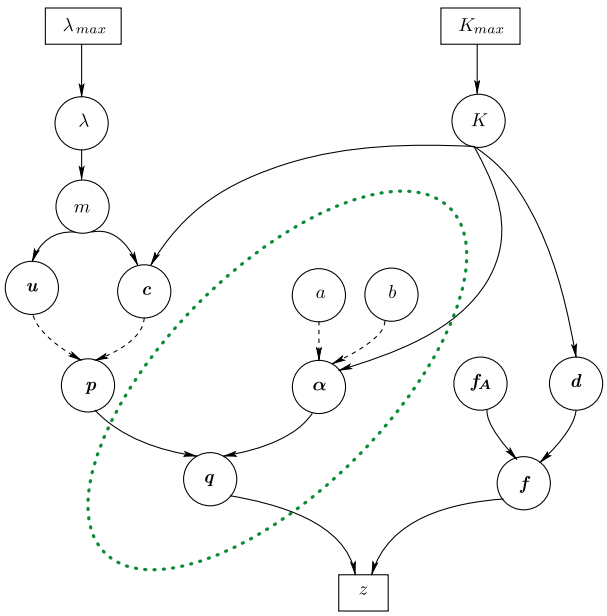

Fig. 5 Directed acyclic graph of proposed model. Continuous lines represent stochastic dependencies, and dashed lines represent deterministic dependencies. Squared boxes enclose data or fixed hyperparameters, and rounded boxes enclose inferred parameters. The thick dotted line encloses, the part of the model proposed that is novel. The other parts are borrowed to Structure or Geneland.

Our model has the direct advantage over Structure to explicitly model the presence of a hybrid zone and therefore to allow one to estimate its width and the intensity of selection or the age of contact, at least in the case of sigmoid clines. For a discussion of sigmoid versus stepped clines (see Kruuk *et al.* 1999). However, we note that in contrast to Structure that explicitly models admixture linkage disequilibrium (Falush *et al.* 2003), our model assumes independence among loci. Associations (LD) between loci under selection lead to a different class of clinal model—stepped clines—not considered here (see Kruuk *et al.* 1999).

Durand *et al.* (2009) proposed an admixture model also based on spatially varying admixture coefficients involving the Dirichlet distribution. It is a general-purpose model that can be justified whenever spatial structure of admixture coefficient is expected. However, it is not specifically tailored for the study of hybrid zones (even though it has been presented in the context). Indeed, their approach does not explicitly model the presence of a contact zone, or to use a mathematical phrasing, their model does not account for the existence of a singularity in space (the contact zone) of genetic variation. What makes the potential usefulness of their approach for the study of hybrid zones is its extreme flexibility, but it does not offer a straightforward way to estimate the width of the hybrid zone and the intensity of selection.

A second salient difference between our approach and that of Durand *et al.* (2009) is the inference machinery. We try to rely as much as possible on Bayesian estimators and therefore on MCMC, including for the estimation of the number of clusters (admittedly with a degree of approximation here) while they resort to likelihood or penalized likelihood methods. In this respect, the initial version of the TESS program (Chen *et al.* 2007) suffered from a number of flaws pointed out by Guillot (2009a,b). Even if the updated model of Durand *et al.* (2009) is an improvement in many respects over Chen *et al.* (2007), it still has some limitations. An obvious one is the impossibility to compare the scenario $K = 1$ against $K > 1$, which makes it impossible to test the null hypothesis of absence of structure. A recent study by Safner *et al.* (2011) suggests also that the new admixture model of Durand *et al.* (2009) may be less accurate than the old no-admixture model of Chen *et al.* (2007).

Our method allows evolutionists to make inference about the location and shape of hybrid zones. It should prove useful in particular in the case of secondary contact between weakly differentiated populations. However, as a final note, we stress that the spatial regression of admixture proportions does not capture all the complexity of hybrid zones: their semi-permeable nature, the fine-scale discordance of clines, the interplay of various component of reproductive isolation etc. Admixture proportions and cline width are only a rough summary of how genomes intermix in hybrid zones, and hybrid zones cannot simply be summarized by logistic variation of admixture proportions. We think the present model will be of great help as a complementary procedure to estimate the

course of hybrid zone centres and selection acting, at least in the case of sigmoid clines. However, we also believe that it will not substitute for detailed analyses of cline shapes and departure from Hardy–Weinberg or linkage disequilibria traditionally conducted in hybrid zones.

Acknowledgements

This manuscript benefited greatly from review comments of Stuart J.E. Baird, two anonymous reviewers and the Subject Editor. This work was partially supported by a grant of the Danish Centre for Scientific Computing and a grant of Agence Nationale de la Recherche (ANR-09-BLAN-0145-01).

References

Barton N, Baird S (1998) Analyse. 1.1. Available at: http://helios.bto.ed.ac.uk/evolgen/Mac/Analyse/.

Bazykin A (1969) Hypothetical mechanism of speciation. *Evolution*, **23**, 685–687.

Bridle JR, Baird SJE, Butlin RK (2001) Spatial structure and habitat variation in a grasshopper hybrid zone. *Evolution*, **55**, 1832–1843.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Cramer J (2003) The origins and development of the *logit* model. *Logit Models From Economics and Other Fields*, Chap. Cambridge University Press, Cambridge.

Durand E, Jay F, Gaggiotti O, François O (2009) Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, **26**, 1963–1973.

Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Guillot G (2009a) On the inference of spatial structure from population genetics data. *Bioinformatics*, **25**, 1796–1801.

Guillot G (2009b) Response to comment on 'On the inference of spatial structure from population genetics data'. *Bioinformatics*, **25**, 1805–1806.

Guillot G, Estoup A, Mortier F, Cosson J (2005) A spatial statistical model for landscape genetics. *Genetics*, **170**, 1261–1280.

Guillot G, Leblois R, Coulon A, Frantz A (2009) Statistical methods in spatial genetics. *Molecular Ecology*, **18**, 4734–4756.

Haldane JBS (1948) The theory of a cline. *Journal of Genetics*, **48**, 277–284.

Harrison R (1990) Hybrid zones: windows on evolutionary processes. *Oxford Surveys in Evolutionary Biology*, Vol. 7, Chap, pp. 70–128. Oxford University Press, Oxford.

Kruuk LEB, Baird SJE, Gale KS, Barton NH (1999) A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153**, 1959–1971.

MacCallum CJ, Nurnberger B, Barton NH, Szymura JM (1998) Habitat preference in the *Bombina* hybrid zone in croatia. *Evolution*, **52**, 227–239.

Macholán M, Baird SJE, Dufková P, Munclinger P, Bímová BV, Piálek J (2011) Assessing multilocus introgression patterns: a case study on the mouse X chromosome in central Europe. *Evolution*, **65**, 1428–1446.

Macholán M, Baird SJE, Munclinger P, Dufková P, Bímová BV, Piálek J (2008) Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone?. *BMC Evolutionary Biology*, **8**, 271, doi:10.1186/1471-2148-8-271.

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Safner T, Miller M, McRae B, Fortin M, Manel S (2011) Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *International Journal of Molecular Sciences*, **12**, 865–889.

Appendix

Inference algorithm

The novel part of the model involves three blocks of parameters: the matrix of admixture proportions $q = (q_{ik})$ and the vector of parameters (a, b). An exact Bayesian inference would estimate them by joint MCMC simulation of (q, a, b) together with any other parameters involved in the model (number of gene pools, tessellation parameters and allele frequencies). We believe that the implementation of this strategy would offer a number of numerical challenges, caused by the joint estimation of q and the number of clusters.

For this reason, we implement an alternative approximate two-stage strategy: first we estimate allele frequencies and cluster locations under the nonadmixture model of Geneland. In a second step, we estimate (q, a, b) by MCMC simulation from the distribution of (q, a, b) conditioned by the data and the parameters estimates obtained from the nonadmixture Geneland run.

Updates of q

We perform updates of $q_{i.}$ into $q^*_{i.}$ where $q^*_{i.}$ is obtained by perturbating two randomly chosen components, i.e. $q^*_{ik_1} = q_{ik_1} + \delta$ and $q^*_{ik_2} = q_{ik_2} - \delta$. When δ is sampled from a symmetric distribution, the Metropolis–Hastings ratio is

$$R = \frac{\pi(z|q^*, \ldots)}{\pi(z|q, \ldots)} \frac{\pi(q^*|\alpha)}{\pi(q|\alpha)} \qquad (eqn\,8)$$

The function $\pi(z \mid q,\ldots)$ refers to the full conditional distribution of the data. The function $\pi(q \mid \alpha)$ is a product of Dirichlet densities.

Updates of a and b

We perform Metropolis–Hastings updates of a. With a symmetric proposal, the acceptance ratio is

$$R = \frac{\pi(q|\alpha^*)}{\pi(q|\alpha)} \frac{\pi(a^*)}{\pi(a)} \qquad (eqn\,9)$$

where $\alpha^*_{ik} = a^* \exp(-d_{ik}/b)$. We proceed similarly to update b.