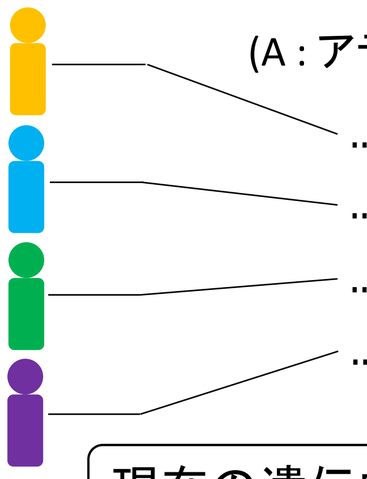
2012年6月15日　統計数理研究所　オープンハウス

Kernel Approximate Bayesian Computation for Population Genetics

中込　滋樹　　リスク解析戦略研究センター　外来研究員

【背景】



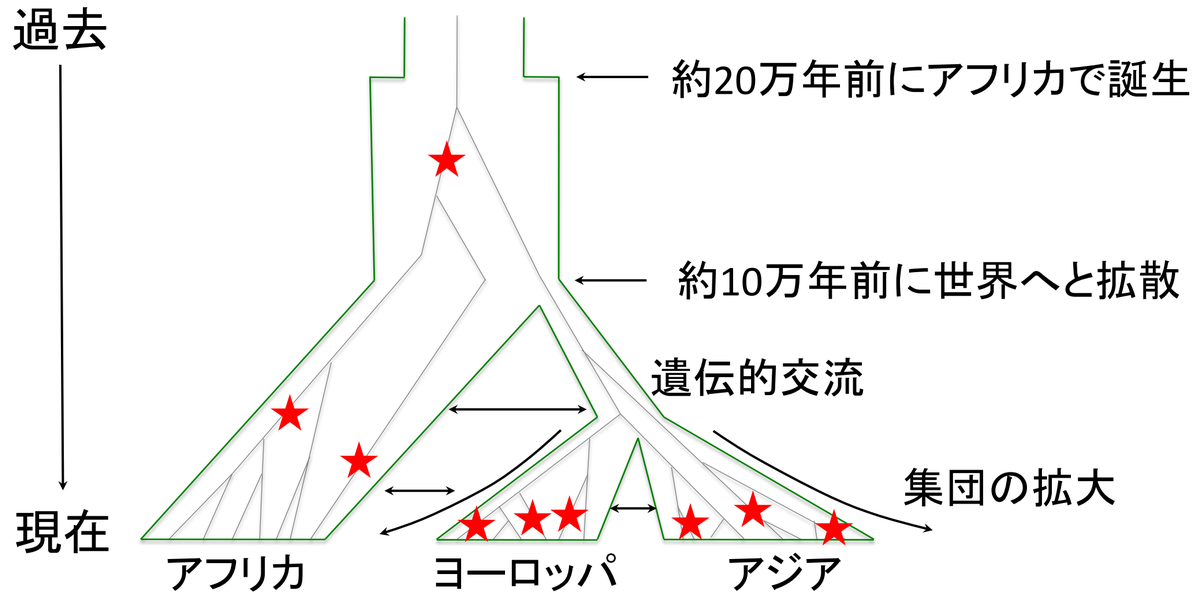
DNA配列
(A : アデニン; C : シトシン; G : グアニン; T : チミン)

...ATTGGCCATAACCCCCGCTT...
...ACTGGCCATAACCCCCGATT...
...ACTGGCCATATCCCCCGATT...
...ATTGGCCTTAACGCCCGATT...

★ 突然変異

現在の遺伝的多様性から過去の歴史を推定する

人類集団の進化モデル



【方法】

ベイズ推論　$\pi(\boldsymbol{\theta}|\mathcal{D}) \propto f(\mathcal{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$
$\mathcal{D}$: data, $\boldsymbol{\theta}$: parameters

● 尤度関数 $f(\mathcal{D}|\boldsymbol{\theta})$が既知の場合

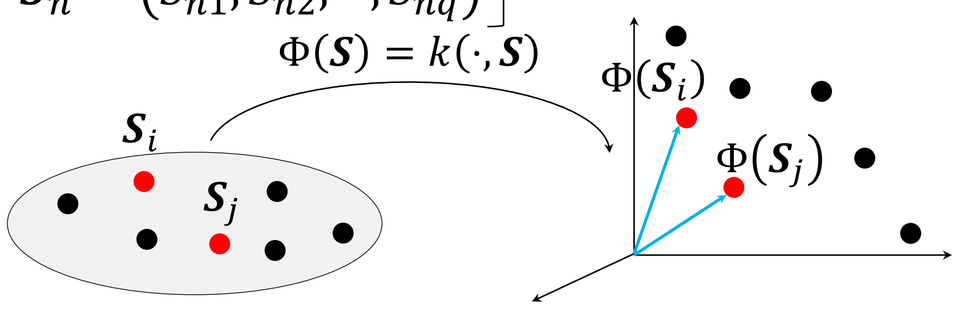
[A] Rejection-sampling method (Ripley 1987)
A1. Generate $\boldsymbol{\theta}_i$ from $\pi(\cdot)$.
A2. Accept $\boldsymbol{\theta}_i$ with probability $f(\mathcal{D}|\boldsymbol{\theta}_i)$,
and go to A1.

● 尤度関数 $f(\mathcal{D}|\boldsymbol{\theta})$が未知の場合

[B] Approximate Bayesian Computation (ABC) method
(Beaumont *et al*. 2002)
B1. Generate $\boldsymbol{\theta}_i$ from $\pi(\cdot)$.
B2. Simulate data $\mathcal{D}_i$ by the model using $\boldsymbol{\theta}_i$.
B3. Accept $\boldsymbol{\theta}'$ if$\mathcal{D} = \mathcal{D}_i$, and go to B1.
→ B3'. Accept $\boldsymbol{\theta}_i$ if $d(\boldsymbol{s}, \boldsymbol{s}_i) < \delta$.
$\boldsymbol{s}$: summary statistics for $\mathcal{D}$
d : metric to measure dissimilarity between $\boldsymbol{s}$ and $\boldsymbol{s_i}$

カーネル法を用いたABC methodの改良
Nakagome *et al*. (2012)

$\boldsymbol{S} = (S_1, S_2, \cdots, S_q)$] Observed data
$\boldsymbol{S}_1 = (S_{11}, S_{12}, \cdots, S_{1q})$
...
$\boldsymbol{S}_n = (S_{n1}, S_{n2}, \cdots, S_{nq})$] Simulated data (n)



Ω : space of summary statistics　$\mathcal{H}_{\boldsymbol{S}}$: reproducing kernel Hilbert space

The inner product between mappings is given by
$$\langle \Phi(\boldsymbol{S}_i), \Phi(\boldsymbol{S}_j) \rangle = k(\boldsymbol{S}_i, \boldsymbol{S}_j)$$

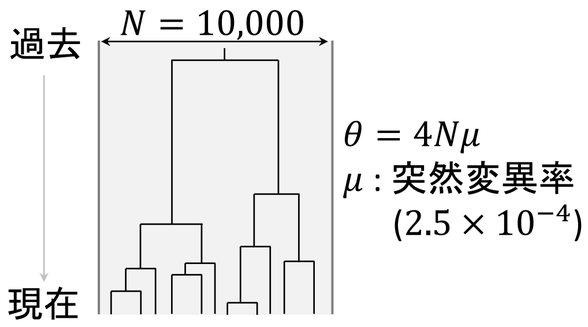
[C] Kernel ABC method
C1. Generate $\boldsymbol{\theta}_i$from $\pi(\cdot)$.
C2. Simulate data $\mathcal{D}_i$ by the model using $\boldsymbol{\theta}_i$.
C3. Compute $\boldsymbol{s}_i$ for $\mathcal{D}_i$, and return to C1.
] $\{(\boldsymbol{\theta}_i, \boldsymbol{s}_i)\}_{i=1}^n$

Kernel Bayes Rule (Fukumizu *et al*. 2011)

The empirical estimator of the kernel posterior mean is given by
$$\hat{m}_{\boldsymbol{\theta}|\boldsymbol{s}} = \sum_{i=1}^n w_i k(\cdot, \boldsymbol{\theta}_i),$$
The weighted coefficient is given by
$$w_i(\boldsymbol{S}) = \sum_{i=1}^n (G_{\boldsymbol{S}} + n\varepsilon_n I_n)^{-1}{}_{ij} k(\boldsymbol{s}_j, \boldsymbol{s}),$$
$G_{\boldsymbol{S}}$: Gram matrix $\left(k(\boldsymbol{s}_i, \boldsymbol{s}_j)\right)_{i,j=1}^n$

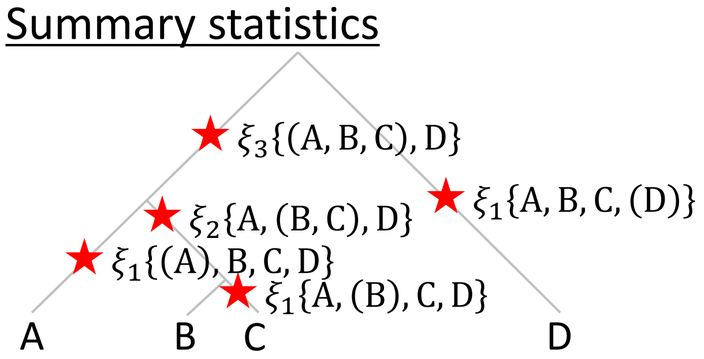
An estimator of posterior expectation of a function $f(\boldsymbol{\theta})$ is given as
$$E[f(\boldsymbol{\theta})|\boldsymbol{s}] = \langle f(\cdot), \hat{m}_{\boldsymbol{\theta}|\boldsymbol{s}} \rangle_{\mathcal{H}_{\boldsymbol{S}}} = \sum_{i=1}^n w_i f(\boldsymbol{\theta}_i).$$

Evolutionary model



$\theta = 4N\mu$
μ : 突然変異率
(2.5×10^{-4})

Summary statistics



ξ_i : the number of segregating sites in which the mutations is present on i sequences.

1. The number of segregating sites (S_{Seg})
→ $\sum_{i=1}^3 \xi_i = 5$
2. Site frequency spectrum ($\boldsymbol{S}_{\mathrm{SFS}}$)
→ $(\xi_1, \xi_2, \xi_3) = (3,2,1)$

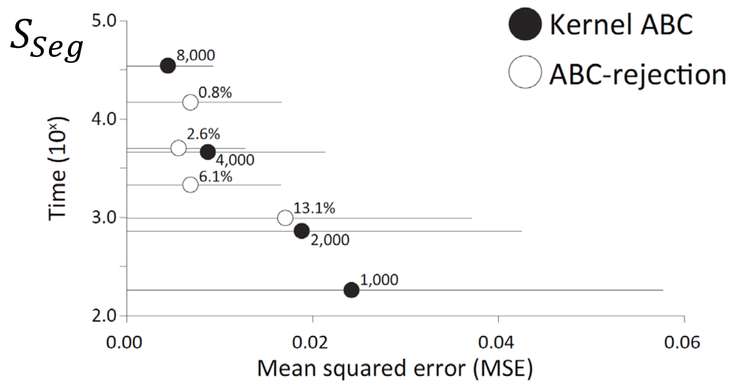
【結果】

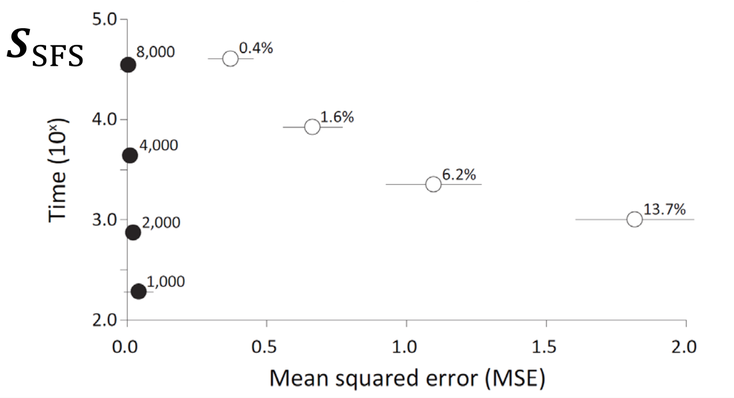
- $f(\mathcal{D}|\theta)$ is computed by importance sampling (Griffiths, 2007),
- Gaussian RBF kernel
$k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$,
(σ : median of pairwise Euclidean distances)

Table 1. Comparison of posterior estimates of θ given $\mathcal{D}$, $\boldsymbol{S}_{\mathrm{SFS}}$, and S_{Seg}.

	$m_{\theta\|\mathcal{D}}{}^{*}$	$\hat{m}_{\theta\|s_{\mathrm{SFS}}}{}^{**}$	$\hat{m}_{\theta\|s_{Seg}}{}^{**}$
Mean	10.498	10.510	9.677
S.D.	0.067	0.044	0.041

*The posterior mean given $\mathcal{D}$ is generated by the rejection-sampling method.
**The kernel posterior means are obtained from 16,000 simulated samples.





【引用文献】

- Ripley, B. D. (1987). Stochastic simulation. *John Wiley & Sons*, New York.
- Beaumont, M. A. *et al*. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025-2035.
- Nakagome, S. *et al*. (2012) Kernel Approximate Bayesian Computation for Population Genetics. arXiv: 1205.3246.
- Fukumizu, K. *et al*. (2011). Kernel Bayes' rule. *Advances in Neural Information Processing Systems 24*, 1549-1557.
- Griffiths, R. C. (2007). GENETREE version 9.0 http://www.stats.ox.ac.uk/~griff/software.html.