Bayesian Clustering using Hidden Random Markov Fields in Spatial Genetics

Olivier François

TIMC (TIMB: Department of Mathematical Biology) - Grenoble

Joint work with

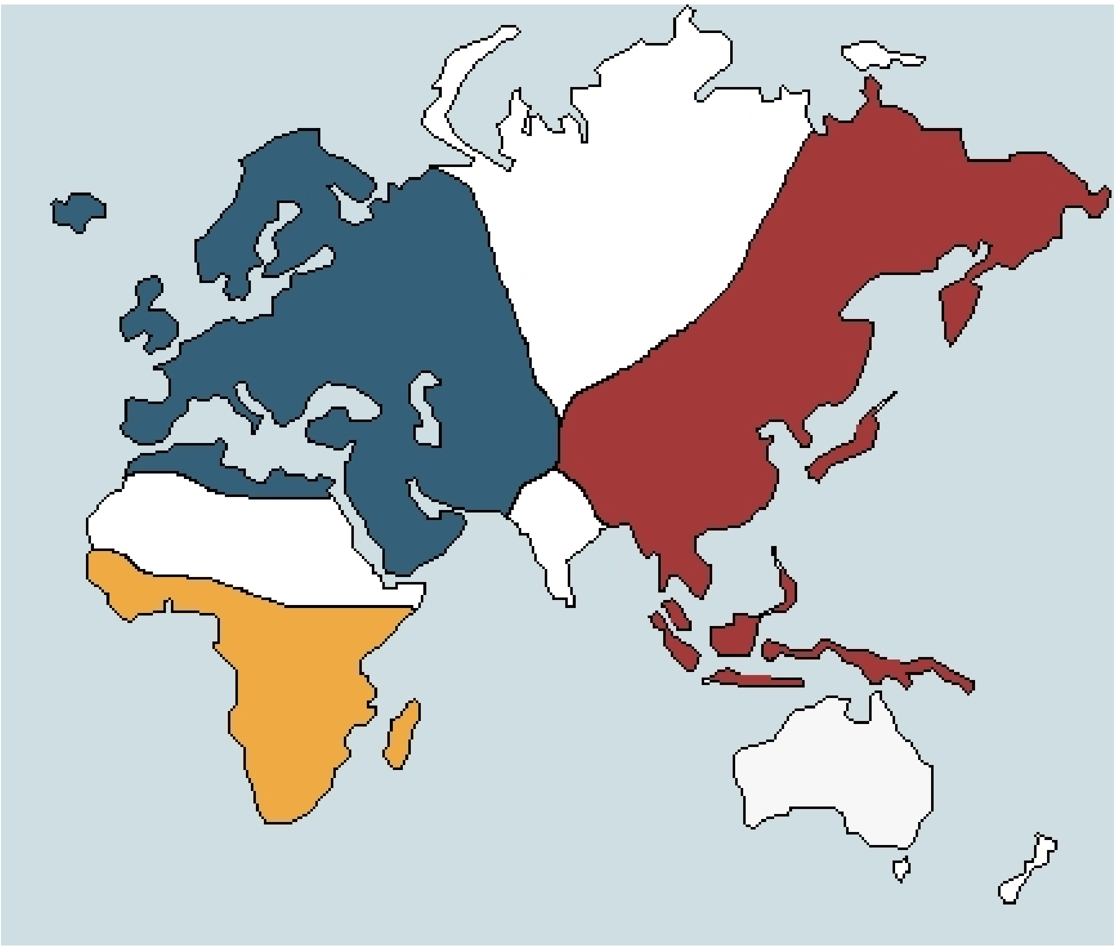
Sophie Ancelet and Gilles Guillot (Engref)

Outline

- Spatial genetics
- Model-based Bayesian clustering algorithms (MCMC)
- New perspective: Hidden Markov Random Fields (HMRF)
- Genetic structure of Scandinavian brown bears

Spatial genetics

- **Statistical genetics**: Use of DNA samples to infer the evolutionary processes that shaped the molecules
- **Spatial genetics**: Explain the spatial variation of DNA among individuals within a population.



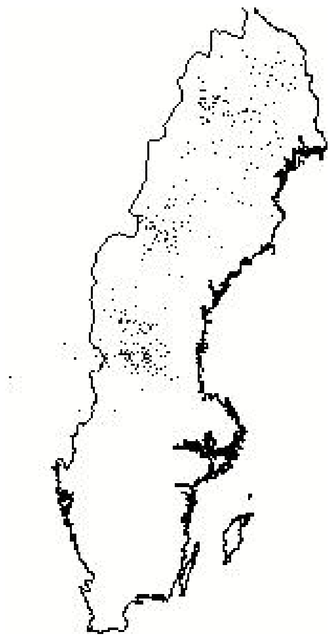
Why is it important?

- Detect the presence of genetically clustered subpopulations (populations are usually defined from subjective criteria)
- Detect changes in population structure: e.g., recent migrations or admixtures
- Issues: Undetected structure may
- lead to conclude that genes are under selection while they are not (low heterozygosity)
- modify Linkage Desequilibrium (correlation among genes) and create wrong associations (of genes to diseases for example)

The data: multilocus genotypes and sampling locations

- Individuals sampled at several geographical sites
- DNA genotyping: each individual genome DNA is amplified at specific loci
- Molecular markers: Short Tandems Repeats in DNA (microsatellites), Single Nucleotide Polymorphisms are the alleles at these loci

acgtagcat||gata||gata||gata||gata||gagatcga



Allele frequencies: the Hardy-Weindberg Law

- Allele frequencies are under equilibrium and remain constant over successive generations
- A consequence of Mendel's law that assumes a panmitic (neutral) rule of mating .

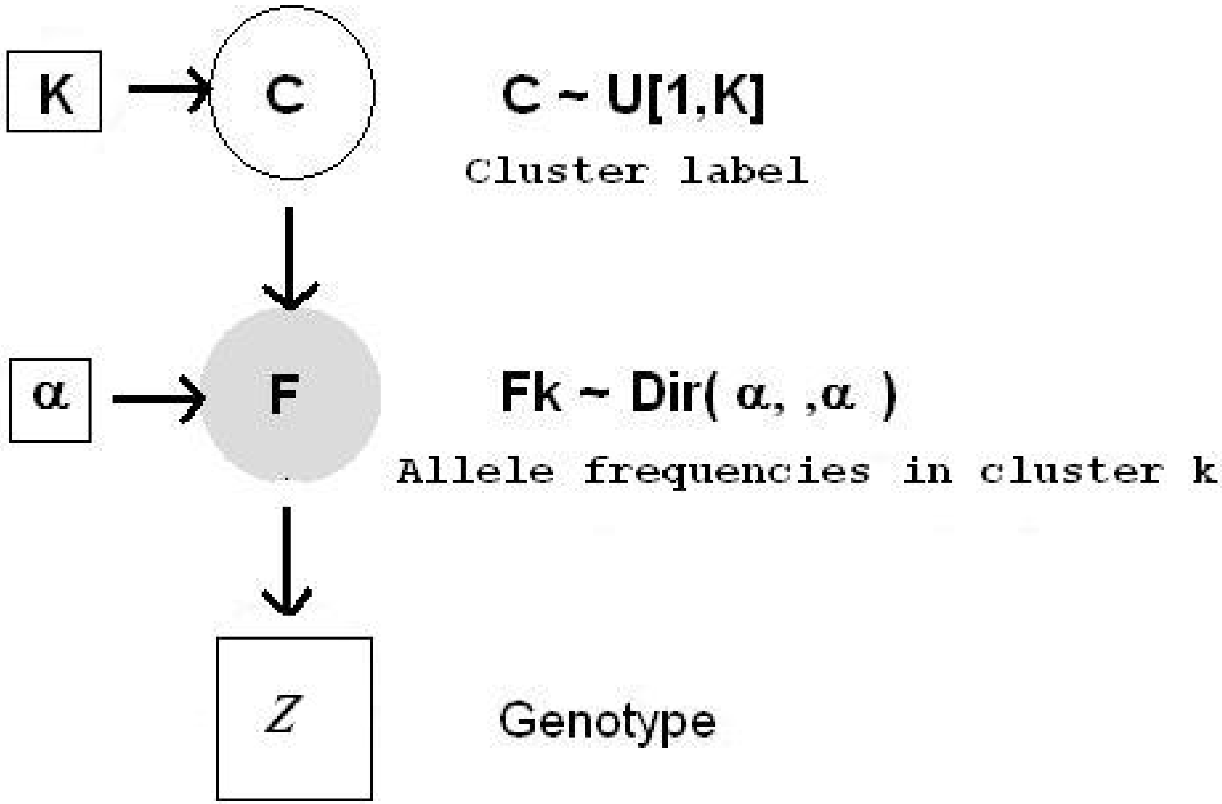
	a	A	
a	p^2	pq	p
A	pq	q^2	q
	p	q	

A Bayesian clustering model

- Model-based approach (Prichard Stephens $\&$ Donnelly, Genetics, 2000).
- The population is subdivided into K subpopulations/clusters
- Each individual may have multiple membership to subpopulations (probabilities π_k)
- Each subpopulation evolves under HW equilibrium. The prior distribution of allele frequencies is a Dirichlet distribution.
- The loci evolve under linkage equilibrium (independence of loci).

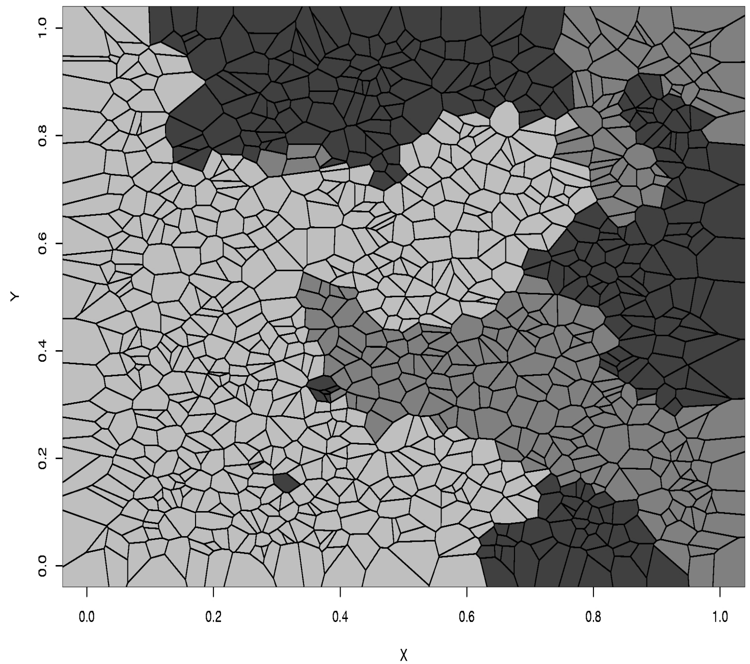
DAG representation

Mixture of Dirichlet + multinomial sampling



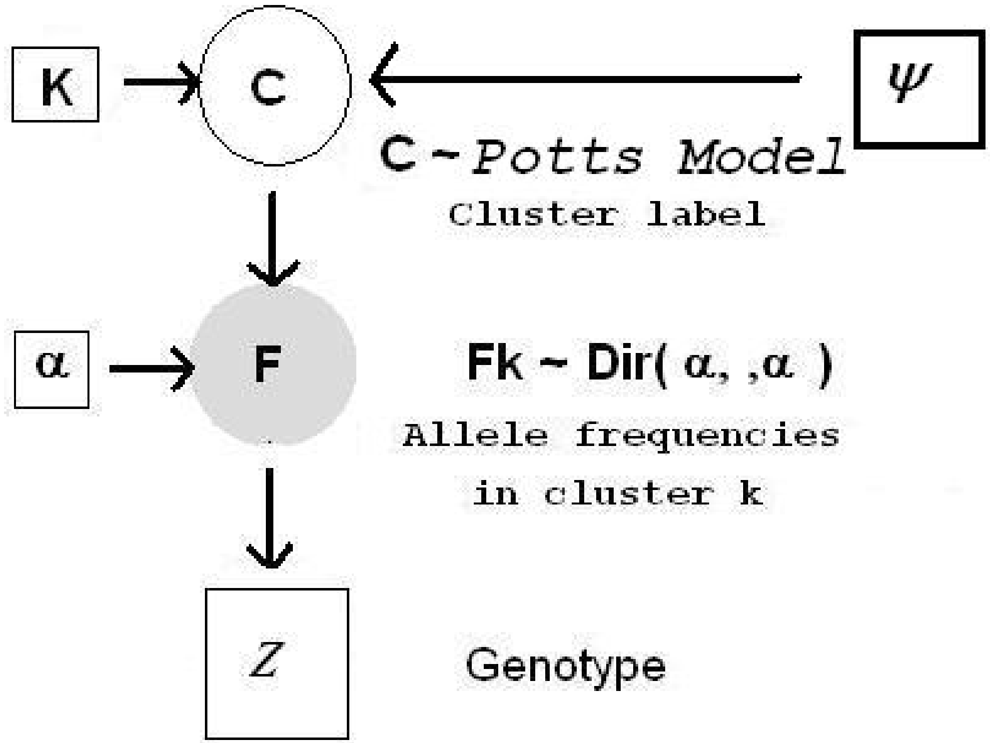
Including spatial priors

- Hidden Markov Random Field: the Potts Model.
- *Individuals living nearby tend to be more alike than those living far apart* (Malécot, 1948; Kimura and Weiss 1964).
- Markov property at the cluster membership level.



New DAG representation

The things to compute: Prob$(C = k|Z = z)$



Model details

- Genotypes: $Z = \{(z_\ell^1, z_\ell^2), \ell = 1, \ldots L\}$, where L is the number of loci and the $z_\ell^i \in \{1, \ldots, J_\ell\}$ are the two copies of the allele at locus ℓ.

- Conditional probability (HW)

$$\mathrm{P}(Z = z \,|\, C = k,\ F = f) = \prod_{\ell=1}^{L} f_{k\ell}(z_\ell^1) f_{k\ell}(z_\ell^2)\,(2 - \delta_{z_\ell^1 z_\ell^2})$$

- The allele frequencies are sampled from Dirichlet distributions (dimension J_ℓ)

$$f_{k\ell}(.) \sim \mathcal{D}(\alpha, \ldots, \alpha),$$

HMRF

- Prior distribution on cluster membership C: MRF for a graph computed from the geographical locations of the sampling sites

$$\mathrm{P}(C_i = c_i \mid C_j = c_j \, , \, j \sim i) \propto \exp\left(\psi \sum_{j \sim i} \chi(c_i, c_j) \right) .$$

- The value $\chi(c_i, c_j)$ represent the interactions between individuals.
- $j \sim i$ means that i et j are neighbours
- Hammersley-Clifford Theorem (1972): representation as a Gibbs measure.

Error Rates in Coassignements - Simulations $K = 2$

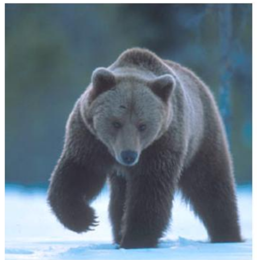
Posterior membership probabilities are computed using a MCMC algorithm.

$F_{\mathrm{ST}} =$ measure of genetic differentiation (low levels $\leq$ 0.05)

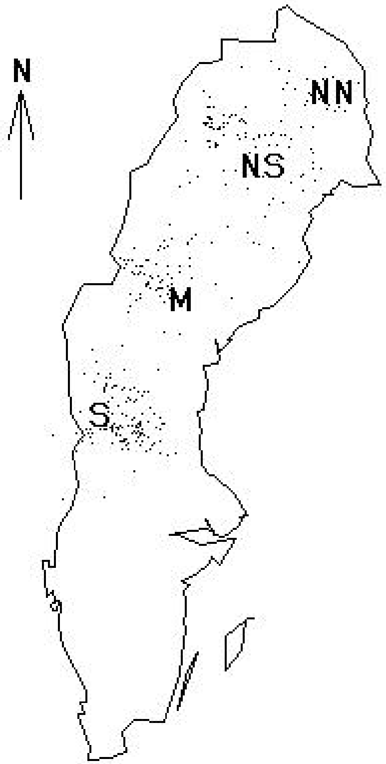
Genet. structure F_{ST}	NON-SPATIAL MODEL	HMRF MODEL	GENELAND
all	16.1	0.7	3.2
$F_{\mathrm{ST}} \leq$ 0.08	26.3	1.6	6.6
0.08$<F_{\mathrm{ST}}\leq$0.09	7.6	0.6	1.4
0.09$<F_{\mathrm{ST}}\leq$0.1	8	0.6	1.4
$F_{\mathrm{ST}} >$ 0.1	8.3	0.2	1.1

Data analysis: Scandinavian brown bears

- 366 brown bears genotyped at 19 microsatellite loci (J. Swenson, Agricultural Univ. Norway), Waits et al. (2001)
- Biologists believed that the population was subdivided into 4 subpopulations (4 areas)
- Areas identified from hunting data during the years 1981-1993 and from the history of the bottleneck

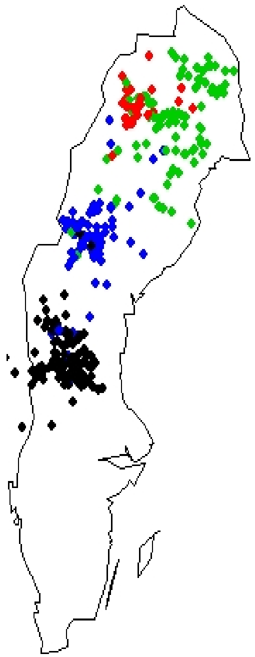


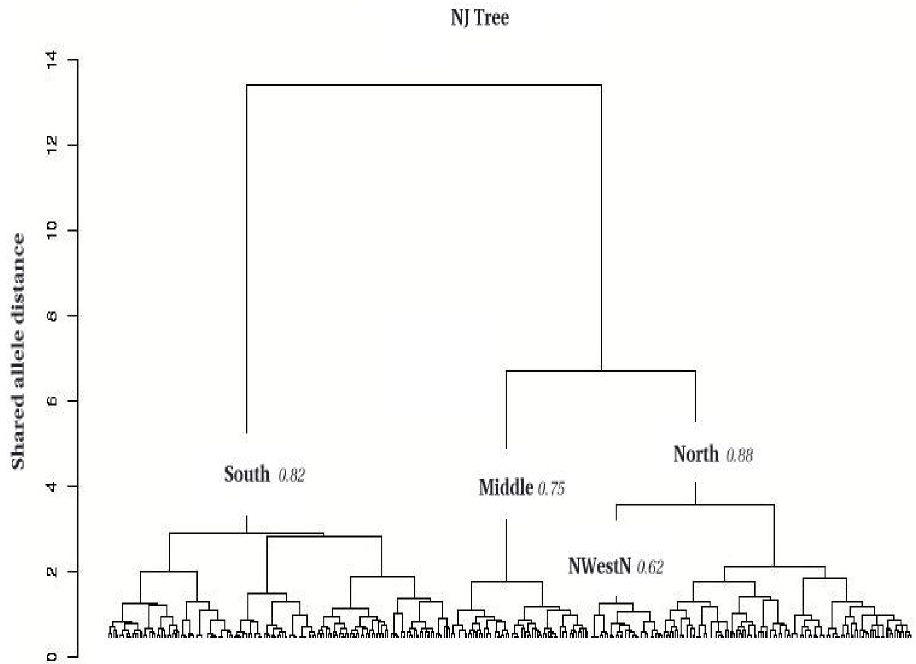
The four predefined subpopulations



Clustering using the HMRF model

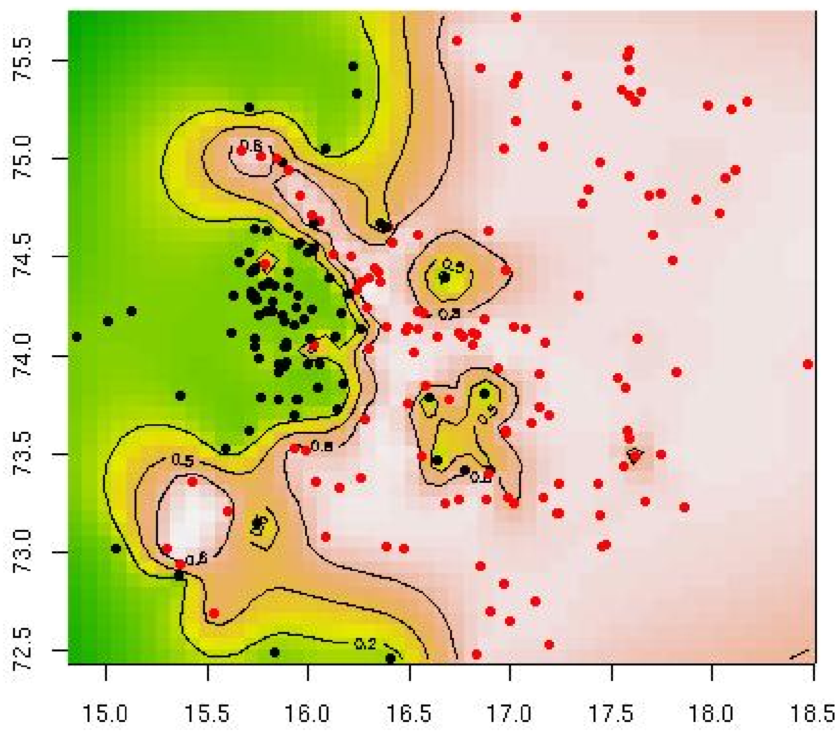
Confirmed by a genealogical method





The Northern NWN cluster

Spatial interpolation of the cluster membership probability, and the posterior assignments to the NWN cluster (black color)



Discussion: The HMRF model

- Choice of K: Bayesian regularisation (cf ridge regression, lasso estimators).
- The log-likelihood writes as

$$L(z, f, c) = L_{\text{non spatial}}(z, f, c) + \psi U(c)$$

  where ψ is the interaction intensity parameter, and $U(c)$ the Energy of a cluster configuration in the Potts model.

- $\psi =$ Lagrange multiplier in a constraint optimization problem where the non-spatial likelihood is optimized while the algorithm attempts to assign a maximal number of neighbours pairs to a same cluster.
- MCMC implantation: extension to include local departures from the HW equilibrium (inbreeding)

Discussion: Bears

- The HMRF hypothesis (Potts) is reasonable because the strong phylopatry of females tends to induce a continuous distribution of genotypes across space
- 2 cluster matches with two predefined populations (S and M)
- But two others don't!
- The NWN (fourth) cluster can be explained by the matriarchal structure of the population.
- Actually, one male was responsible for 88% of the descendants in the group, the male was the father of 70% of them, grandfather of 12% and great-grandfather for 6% of them, and probably the uncle for 9% of them (parentage analysis).
- Conclusion for the bear conservation policy: No reasons for distinguishing the NS and NN regions.