Some background on the program ∂a∂i

Desiree Harpel and Peter Beerli

The program ∂a∂i estimates complex population models using the allele frequency spectrum (AFS) of a sample of individuals from one or several location. The methods needs a large amount of data. It is based on an diffusion approximation of the population genetic model.

I. Allele frequency spectrum

Lets assume we have this data (Table I, left panel). We can calculate the site frequency spectrum for the whole population (1-D spectrum) or we can calculate the joint frequency spectrum (2-D, Table I, right panel), with n populations this will be come more and more cumbersome because we will need to fill in values into an n-dimensional hypercube.

Without many markers the site frequency spectrum is rather empty and non-informative, but with with large contiguous genomic data the method becomes powerful. The site frequency spectrum needs to be oriented, if we have an ancestral sequence or an outgroup then we can know the ancestral alleles at each locus, then the site frequency spectrum records the frequency of the derived allele. Often we do not know the ancestral allele, in these cases ∂a∂i allows to use the folded AFS where it is assumed that the rare allele at a particular locus is the derived allele.

II. Markov Process

We can think of a process $f_j(t)$ that generates the expected number of loci at which a derived allele is found on chromosome j at time t where we have $1 \leq j \leq 2N$ where N is the number if diploid individuals. We expect that with a mutation process and with a propagation of genetic material through offspring that the state will eventuall change, this can be expressed in a recurrence equation

$$f_j(t+1) = \sum_{i=1}^{2N} P(j|i) f_i(t) + \mu_j(t) \quad \text{with} \quad 1 \leq j \leq 2N \tag{1}$$

then explains the change of the states through time, we assume that μ is the mutation process. When we assume that the populations behaves like a Wright-Fisher population then

$$P(j|i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \tag{2}$$

This process does not have any memory; it only remembers the last state and uses that to generate a new state; such processes form Markov chains. These recurrence equations are often simplified using a diffusion approximation.

III. Diffusion approximation

A common theme is to take the recurrence equation and assume that the change is independent of the t and that the change is drawn from a Normal distribution with some standard deviation, we can then write

$$X(t+\Delta t) = X(t) + \Delta X \tag{3}$$

$$\Delta X \sim M(X,t) \tag{4}$$

with variance $\sigma(X,t)$

$$X(t+\Delta t) = X(t) + M(X,t)\Delta t + \sigma(X,t)\epsilon\sqrt{\Delta t} \tag{5}$$

with ϵ as the error of a standard normal; this leads to a new expression of ΔX, if we make this very small it becomes the differential

$$dX^a = M(X,t)dt + \sum^{K} \sigma^{ab}(X,t)dW^b \tag{6}$$

which leads to

$$\frac{dX^a}{dt} = M(X,t) + \sum^{K} \sigma^{ab}(X,t)\frac{dW^b}{dt} \tag{7}$$

Applying now this recipe to our population genetics framework we get the master formula ∂a∂i uses:

$$\begin{aligned}\frac{\partial\phi}{\partial\tau} =& \frac{1}{2}\sum_{i=1}^{P} \frac{\partial^2}{\partial^2 x_i}\frac{x_i(1-x_i)}{v_i}\phi \\ &- \sum_{i=1}^{P}\frac{\partial}{\partial x_i}\left(\gamma_I x_i(1-x_i) + \sum_{i=1}^{P} M_{i\leftarrow j}(x_j - x_i)\phi\right)\end{aligned} \tag{8}$$

where ϕ is the AFS augmented by the time $\phi(x_1, x_2, ..., x_P, t)$.

This partial derivate equation can be solved numerically for a particular population model. The Formula 8 shows parameters for the effective population size $v_i = N_i/N_{\text{ref}}$, selection coefficient $\gamma_i = s_i/N_{\text{ref}}$, and migration $M_{i\leftarrow j} = N_{\text{ref}} m_{i\leftarrow j}$. The python module *dadi* has functions to handle population growth, immigration, population size and selection. Thus one can assemble bottleneck models with growth with divergence and migration: very complex models indeed.

The PDE is solved by finite differences method, this is considered a crude method by scientific computing standards but usually works well. The key for the analysis is not solving the PDE but comparing the expected site frequency spectrum given the population model with the observed site frequency spectrum.

Table I
An example single nucleotide polymorphism data set and the site frequency spectra. Left: raw data. Middle: 1-D site frequency spectrum. Right: 2-D Joint frequency spectrumThe Occurrences record how often a particular site pattern was seen in the data. The Pattern reflects a particular pattern, such as 1 means 1 chromosome is different at a particular sites from all others, 2 means that two are different. The numbers reflect the derived allele if an outgroup is available, otherwise it is the rarer allele.

Individual name	SNPs
0-1	ATAGACG
0-2	ACGGACG
0-3	ACAGAAG
0-4	GCGGACC
1-1	GCATTCC
1-2	GTATTCC
1-3	GCATACC
1-4	ACATACC

Patterns	Occurences
0	0
1	1
2	3
3	1
4	2
5	0
6	0
7	0
8	0

	Patterns in Population 2								
Patterns in Pop 1	0	1	2	3	4	5	6	7	0
0	1	1	0	1	0	0	0	0	0
1	0	1	0	1	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0

IV. Likelihood calculation

To simplify the analysis Gutenkunst et al. [2009] assumes that all loci in the spectrum are independent, this allows to calculate a likelihood where we assume that every entry in the AFS represents a success. This allows to phrase the likelihoods in terms of Poisson variables. So for example the probability of seeing 5 successes with a mean success of 3 is

$$P(x;\mu) = \frac{e^{-\mu}\mu^x}{x!} = \frac{e^{-3}3^{-5}}{5!} = 0.100819 \tag{9}$$

replacing x with the observed site frequency spectrum $S(d_1, ..., d_P)$ and also replacing μ with the expected frequency spectrum using the population model M we can write the likelihood for each part and the total likelihood $P(S|\Theta)$

$$P(S(d_1, ..., d_P); M(d_1, ..., d_P)) = \frac{e^{-M(d_1,...,d_P)}M(d_1, ..., d_P)^{S(d_1,...d_P)}}{S(d_1, ...d_P)!} \tag{10}$$

$$P(S|\Theta) = \prod_{i=1}^{P}\prod_{d_i=0}^{n_i} P(S(d_1, ...d_P)|M(d_1, ..., d_P)) \tag{11}$$

This leaves now to explain $M(d_1, ..., d_P)$ which is simply taking the average of all AFS values calculated by the PDE for a particular set of parameters Θ, M is calculated as the integral over all entries in the site frequency spectrum [details see in Gutenkunst et al., 2009]. This $P(S|\Theta)$ (or equivalently written $L(\Theta|S)$) is a 'Composite' likelihood because we assume independence between loci despite knowing that this is not correct. We treat each SNP as locus this allows to simply multiply the probabilities of all the sites without considering the covariance structure among the loci. This pretends that each locus has full information whereas linkage disequilibrium will force loci to covary: two loci that are completely linked have the same information as one. Usually composite likelihoods are much faster and easier to calculate than the correct likelihood, but still deliver the same or similar maximum of the function to optimize. Composite likelihood falls short to describe the confidence interval correctly and delivers commonly to narrow confidence limits.

V. Confidence intervals

Confidence intervals can be calculated in three different ways with ∂a∂i. The Fisher information (using the second derivative) assumes that all loci are independent. Use of the Godambe matrix that can take into account correlations, and parametric boostrap. The parametric bootstrap uses the simulation program MS [Hudson, 2002] to generate samples using the population model parameters Θ. ∂a∂i has functions that can run many of these simulations and convert the results into site frequency spectra that than can be used to generate a Null distribution, the observed distribution should be in the center and not the tails of this distribution to suggest that the parameters represent a good interpretation of the data.

Bibliography

Ryan N. Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, 5(10):e1000695, October 2009.

R R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

VI. Disclaimer