Laboratory excercise written by Andrew J. Crawford <andrew@dna.ac>
with the support of CIES Fulbright Program and Fulbright Colombia.
Enjoy!

Filogeografía

BIOL 4211, Universidad de Los Andes, Bogotá
25 de enero a 01 de abril de 2005

Lab 4 – IM, Isolation with Migration

18 de febrero 2006

NB: Always download the latest version of all software
and read carefully the latest accompanying documentation.

Today we will be estimating population genetic (in the broad sense) parameters by applying to our data the "isolation with migration" model of Jody Hey, Rasmus Nielsen, and colleagues. The model is applicable to a pair of populations that are each others' closest relatives, or to a pair of sister species that are well-sampled genetically. The two groups (populations) certainly do *not* have to be reciprocally monophyletic, but the combined data set of both populations (groups) should form a monophyletic group. In other words, if possible, investigators should avoid leaving out sequences from the analysis that are genealogically "inside" the group under analysis.

For best results, the two populations or sister species would each be represented by DNA sequence data from multiple genes from 10+ individuals. For most of the lab groups in this course we have available just one mitochondrial gene. For many of the groups, only one or a few DNA sequences were collected from any single locality. Each lab team will have to decide how best to apply an IM model to their data, keeping in mind that the model should be applied to a group of sequences that together form a monophyletic group, e.g., a phylogeographic clade or a pair of sister species. In choosing a pair of populations, you can use both your distribution map for your species as well as the NJ tree you obtained during the *Clustal* lab. Your first choice would be a pair of populations that each have 10+ DNA sequences. Your second option would be to combine samples from nearby localities such that you have two groups of DNA sequences that could form sister clades and each clade contains as many DNA sequences as reasonably possible.

The second important decision for each lab group is to decide what type of population genetic model they want to evaluate. The most complex model contains 7 parameters, but depending on the interests of each lab team, the limitations of the data, and the type of question each group wants to address, the students in each lab group may decide on the most complex model or any limited case (subset) of the full model. Read the introduction to the lab (part **I.** below), decide on one (or more) models to analyze, and decide on one (or more) paired sets of DNA sequences to analyze. Once these two decisions are made, proceed with the lab for today. Obtain (or assume) a mutation rate (part **II.** below), test the software (part **III.** below), prepare your data infile (part **IV.** below), implement your preferred model as a command line (part **V.** below), and finally run the program.

Software

IM version "late 2005" (probably PC version – Mac version may not be functional at this time)
`http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#IM`

Citations

Hey, J. 2005. On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas. *PLoS Biol* 3:e193.

Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. *Genetics* 158:885-96.

Papers are available on the *Filogeografía* webpage of references. Note, *PLoS Biology* is an open access journal freely available online.

Further Information

You should first open a copy of the *IM* documentation "`IMdocumentation.pdf`" that may be found on your computer along with the program, or on Jody Hey's webpage (see above), or on the *sicua* webpage of this course, under Lab 4.

GOALS

- Gain familiarity with the *IM* software.
- Understand the benefits and potential pitfalls (problems) with running Bayesian MCMC analyses.
- Evaluate whether your *IM* analyses are providing reliable results.
- Understand and interpret the output of *IM* in terms of the evolution and demography of your species.
-

I. Introduction to *IM* and Bayesian MCMC analysis

IM is a software package that implements the MCMC method for the analysis of genetic data under the "Isolation with Migration" model of population divergence. The latest version of *IM* now can estimate up to seven parameters of a model that describes the division of one population (ancestor with θ_A) into two populations at time t, after which point in time migration may still occur between the two populations. This migration may be asymmetrical (m_1 and m_2). The two extant populations are described by θ_1 and θ_2. At time t when the ancestral population split, the first daughter population may be described as being founded by a proportion, s, of the ancestral population, and the second population therefore was founded by (1-s) of the ancestral population. Keep in mind that this is the full model, but students may decide to analyze a logical subset of this model.

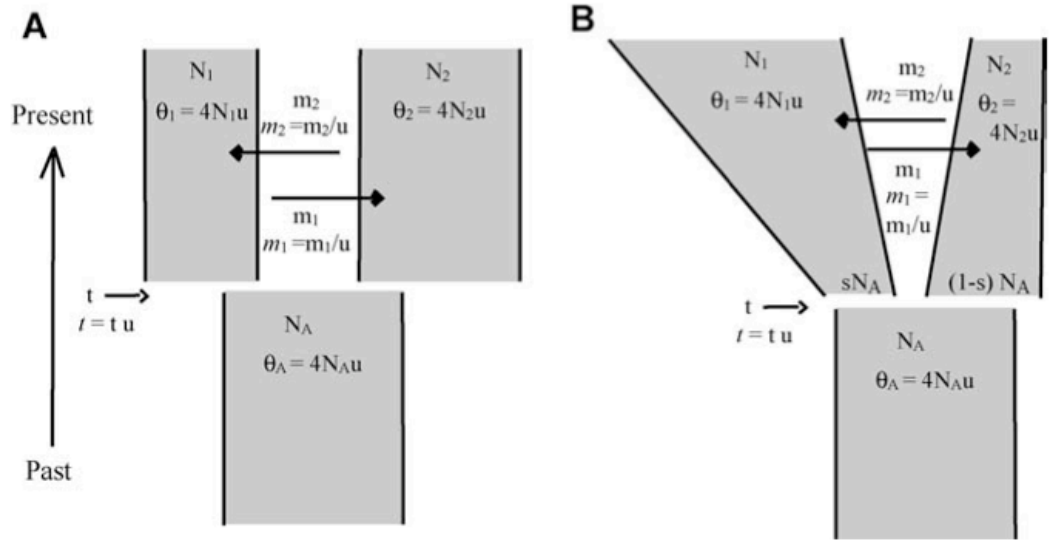

Figure 1. Isolation with Migration Models

Note, in the above figure, the horizontal arrows indicate the direction of migration in coalescent time, *i.e.*, backwards time. In the forward-time sense, m_1 refers to haplotypes moving from population 2 *into* population 1. Students should consider running a limited case of the full model. Subsets of this model may include: symmetrical migration, no migration, no s parameter, or all θ's may be equal. *IM* can also be used to fit a two-island (aka, stepping-stone) equilibrium model (setting t to be very large) or a single, constant size population model ($t = 0$, with only θ_A estimated). If you have multiple independent genes, the model may be made more complex by allowing each locus to have a different migration rate, but this is not recommended for today.

Assumptions of the model: Please keep these in mind (1) There should be no other populations more closely related to the two sample populations than these two are to each other. (2) There should be no other populations exchanging genes with the sampled populations or their ancestor. (3) No evidence of natural selective in your data (evidence of demographic expansion or contraction is fine, of course, because *IM* accommodates change in population size. (4) No recombination within loci. The validity of this latter assumption can be evaluated using the 4-gamete test as implemented in *dnaSP*, but would be of particular concern only for nuclear gene sequences. (5) Free recombination

between loci. Multiple mitochondrial genes should therefore be analyzed as one long gene. (6) One or more of four possible mutation models is applicable to the data: Infinite sites (IS), finite sites (HKY), step-wise mutation model (SMM) for microsatellites, or an IS+SMM compound model. Each locus in the data can have a different model. For our purposes, only IS and HKY models are of interest. If you have multiple substitutions at any single site in your data, you will have to use the HKY model. If you incorrectly assume the IS model *IM* will warn you.

The MCMC method: Briefly, the program tries to estimate the marginal posterior probability distribution for each parameter in the model (probability of the parameter given the data). The program starts with a prior distribution of values for each parameter (probability of the parameter values before seeing the data), and estimates from simulation the marginal likelihood (probability of the data given the parameter) for each of the parameters, integrating over a series of possible genealogies. The most important and difficult part of the analysis is obtaining an adequate representation of the posterior distribution of the parameters. The posterior distribution is estimated from a long run of a Markov chain Monte Carlo (MCMC) analysis. During each step of the run, one or more parameters are updated to new values and any new set of parameter values that is accepted is added to the collection of samples representing the posterior distribution. In each run, the MCMC must first reach the stationary distribution before samples are collected, and each sample should be independent of the one that preceded it, though this is often not the case (see ESS below). Analogous to searching tree-space in phylogenetics, we also want the MCMC chain to adequately sample parameter space and not get stuck in local optima of joint likelihood. We'll talk more about Bayesian MCMC analyses in an upcoming lecture on Bayesian phylogenetics, but for now understand that ideally one would want to run multiple analyses, each for a long duration, to get to a point where one may begin to trust the results. However, today, we have less than 4 hours!

One way to improve mixing of the MCMC run over parameter-space is to add semi-independent chains, and have each succeeding chain explore ever-wider areas of parameter space to add to the sample representing the posterior distribution. Such methods are called Metropolis-coupled MCMC, again more on this in a later lecture on Bayesian phylogenetics.

IM measures the independence of successive and cumulative samples by recording a value called the Effective Sample Size (ESS), which may be regarded as the number of truly independent samples taken to estimate the marginal posterior probability distribution of the values of a given parameter. Normally, ESS values reported by *IM* should be regarded as reliable only after 1 million steps in the MCMC run. But we won't have time today for so many steps.

II. Obtaining a mutation rate for your organism

To estimate either effective population size or the time since divergence of your population or species pair, you will need to assume some value for the mutation rate, μ. Recall that in coalescent theory, time in marked in units of N_e, and for a single population at equilibrium, $\theta = 4N_e\mu$ for a nuclear gene. Although μ is usually reported as "per site, per generation," for *IM* this rate must be in terms of "per gene, per year." Next week we will talk about rates of molecular evolution and the difference between substitution rate and divergence rate. For now, accept that the rate of substitution at silent sites (K_S or d_S) may provide an adequate estimation of the mutation rate, μ. Estimated K_S in vertebrate nuclear genes tend to be between 1×10^{-9} and 5×10^{-9} substitutions per year. For frog mitochondrial DNA, I have estimated K_S values between 14×10^{-9} and 37×10^{-9} substitutions per year. For example, assume that you have a mitochondrial gene in a turkey and this gene experiences a mutation rate of 40×10^{-9} per site per year. If your mitochondrial gene was 500 base pairs (bp) long, your mutation rate per gene would be $500 \times 40 \times 10^{-9}$ or $20{,}000 \times 10^{-9}$ or 2×10^{-5} per gene per year. In *IM* you would report this value as 0.00002 in your input file. Assuming a higher mutation rate (μ) would yield smaller estimates of effective population size (N_e) and younger estimated times since divergence (t). Look again at the figure above.

III. Testing the software

In the folder containing the *IM* program, there should also be a file called "mydata." This file contains an example data set to practice using *IM*. We will use this program to confirm that *IM* is installed correctly. *IM* runs in the DOS window. Find and Launch the Command Prompt application (Símbolo del Sistema). In Windows OS, this application might be under Start > All Programs > Accessories (Inicio > Todos los Programas > Acesorios). Navigate to the directory containing the *IM* software. The following DOS commands are helpful:

dir = display contains of current directory.

cd = change directory (cambio de carpeta), e.g., on my computer after launching Cmd Prompt, I type:

```
cd ..\..\Program Files
```

Note ".." = "one directory up" in the hierarchy.
Launch *IM* and at the command prompt type the following command all on one line:

```
im -i imtestdat.u -o mydata.out  -q1 10 -m1 10 -m2 10
-t 10 -b 10000 -L 0.5 -s 123  -p 7
```

Note, in the above command, there is no lower case "l" ("el"), those are the number 1 ("one"). Hit 'return.' Hopefully the program is now running an analysis. The analysis will last less than an hour, but feel free to quit the program at any time using ctrl-c. "`imtestdat.u`" should be the name of the sample data file, but you should confirm the name and location of this file.

IV. Preparing your data for *IM*

Step one, as always, is "read the documentation," therefore locate and open a copy of "`IMdocumentation.pdf`" if you have not done so already. Someday evolutionary biologists will establish universal standards and protocols for organizing and analyzing molecular data, but for now we often must re-format our data for each new software program we want to use. *IM* is one of the more fussy (less permissive) concerning the format of the input data file. In the documentation, turn to the section entitled "Input File Format" (page 15). Note, the only ambiguous base that *IM* accepts is `N`. Other ambiguities such as `Y`, `W`, `K`, `S`, `R`, should be replaced with `N`. Below is an example data set and some tips on how to re-format your own data to be read by *IM*. Probably you should also remove gapped sites, but *IM* may do this automatically.

```
Example data for IM, with 1 gene, 2 pops, 2 & 1 sample per pop, and 112 bp.
Bogota Anaheim
1
COI 2 1 112 H 0.25 0.00001
Bogota504 ATTAACCTCTACACTATGATACTGATAGCCCTCAGCCTCCTCTTAGGTTCCTCAATTACTATCTCTAG
GAATCACTGAATCTTGGCCTGACTTGGCCTAGGGATTAATACCT
Bogota345 ATTAACCTCTACACTATGATACTGATAGCCCTCAGCCTCCTCTTAGGTTCCTTAATTACTATCTCTAG
GAATCACTGAATCCTGGCCTGACTTGGCCTAGAGATTAATACCT
Anaheim23 ATTAACCTCTACACTATGATACTGATAGCCCTCAGCCTCCTCTTAGGTTCCTCAAATACTATCTCTAG
CAATCACTGAATCCTGGCCTGACTTAGCCTAGAGATTAATACCT
```

line 1: arbitrary text. Type any useful information here.
line 2: the names of your two populations (or species).
line 3: the number of independent genes in your data file. Multiple mitochondrial genes could as one gene!
line 4: name of gene, number of samples in pop. 1, number of samples in pop. 2, length of each DNA sequence, letter indicating mutation model (H = HKY = finite sites), inheritance scalar (0.25 = mtDNA or Y-chromosome), and mutation rate per gene per year. See documentation for more options.

There is a small amount of flexibility in data format, but the safest format is probably to make the first position of the DNA sequence be the 11th character (or space) on each line of data. The easiest way to get your data into this format will be to open your single-gene NEXUS files in *PAUP**, execute the data, and then remove all the sequences except the ones from the first of your two "populations." Be aware, however, that PHYLIP (and *IM*) format allows only very short samples names (10 characters). You may also try using *BioEdit*, *dnaSP*, *MEGA*, *Mesquite*, or *MacClade* to export your data as **NBRF**, which would be preferable to PHYLIP. Or you may simply use a text editor to make your NEXUS file conform to the above example and to the guidelines in the *IM* documentation ("Input File Format," page 15).

The following instructions are for *PAUP**. In *PAUP** the command to delete taxa is:

```
DELETE taxon-list [/ONLY];
```

Export your data as PHYLIP format and save under a new name (using the name of your population). In *PAUP** the command is:

```
EXPORT [options];
```

Luckily the default settings are `format=PHYLIP` and `interleaved=no`. Now restore all the data, and the exclude all the samples except those of your second population (or species). In *PAUP** the command to restore taxa is:

```
UNDELETE taxon-list [/ONLY];
```

However, if you add the "`/ONLY`" option to your `delete` command, this should effectively restore all the taxa not in the corresponding taxon-list of samples to be deleted. Export these data in PHYLIP format and save this file under the name of your second population. Open the first population data file in WordPad or other basic text editor and if your

data are in PHYLIP, you will need to add a <return> before each DNA sequence. If your data are in NBRF, you will need to remove the ">" symbols fro from in front of the name of each sample. Type in the necessary first four lines of the data file (see above and see documentation). Open the data file for the second population, clean up the data, and copy-&-paste the results into the bottom of your first file. Save this new file under a new name, such as MyData_IM_input.txt This file should be in the *IM* folder along with the *IM* application. There should be no lines or text separating the last sequence of the first population from the first sequence of the second population.

If *IM* still won't read your data (see below), you may have to open Jody Hey's example file, save under a new name, and paste in your data without changing the top lines or the bottom line (if possible). I don't know why, but this worked for me. There may be some incompatibilities between software programs (e.g., *IM* vs. *WordPad*) in the structure or content of even simple text files. You can also try the same trick using the example file on the *sicua* webpage for the *Filogeografia* course. If you want this file, look in the folder, Lab 4, and download the file called: `imtest_mt_LCvPNRO.txt`

V. Designing your analysis

You and your lab partner need to decide under which model you want to analyze your data. For example, I would NOT recommend you use the *s* parameter. If your two populations or species are very diverged and geographically far apart, you may be more interested in just the divergence time. Therefore, you may decide to fit your data to a 4-parameter (no *s*) model with no migration, instead of the 6-parameter IM model. Also, the fewer data you have, the fewer parameters you should try to estimate. You may decide to assume that all three θ parameters are equal, or assume symmetrical migration rather than asymmetrical. These options are controlled by the –j option in the command line. See documentation for more information, especially the sections "Command Line Options" and "Explanation of Command Line Terms," pages 21-30.

One your team decides on the analysis, you will have to design your command line. I suggest opening a text file which will contain your command lines that you design and test. Then just simply copy-paste each command as you want to run it. Also, in lab we can only run short or quick analyses, which will not be satisfactory. If you have the opportunity to run longer analyses, please do so in the coming week. Here are some example command lines:

(A.) for mtDNA data for recently diverged frog populations:

```
im -i datafile_name.txt -o datafile_name.out -q1 10 -m1 5 -m2 5 -t 10 -u 1 -b
      50000 -fl -n 2 -g1 0.05 -L 0.5 -p 2357
```

(B.) for nuclear data for recently diverged frog populations):

```
im -i datafile_name.txt -o datafile_name.out -q1 10 -m1 10 -m2 10 -t 5 -u 1 -b
      50000 -fl -n 2 -g1 0.05 -L 0.5 -p 2357
```

(C.) for mtDNA for deeply divergent frog populations, setting migration to zero (Isolation-only model):

```
im -i name_of_datafile.txt -o name_of_datafile.out -q1 5 -m1 0 -m2 0 -t 15 -u 1
      -b 50000 -fl -n 2 -g1 0.05 -L 0.5 -p 2357
```

What the commands mean:

`im` = launch program
`-i datafile_name.txt` = name of input file
`-o datafile_name.out` = name of output file
`-q1 10` = q is a scalar for θ, 10 = set the prior to the interval 0-10 for all 3 θ parameters.
`-m1 5 -m2 5` = sets 5 as the maximum value of the prior for either migration rate.
`-t 10` = sets as 10 the maximum value of the time parameter, $t\mu$, where t = time since divergence.
`-u 1` = sets generation time in number of years to 1. (`-u3` would mean it take 3 years to mature.)
`-b 50000` = sets the 'burn-in' period of the MCMC to 50,000.
`-fl -n 2 -g1 0.05` = Metroplis-coupled MCMC with 2 chains, the second being heated by 0.05.
`-L 0.5` = means the outfile will be written every half hour.
`-p 2357` = 4 additional histograms to be included in the outfile.

VI. Running and Stopping *IM*

Launch the Command Prompt application, and navigate to the directory containing the *IM* software, as above. Confirm that you are in the IM folder and so is your data file. Run the command line you created above.

To stop the program, one should change the name of the file 'IMrun' (in the *IM* folder) to 'STOP_IMrun'. When the next -L = 0.5 hour period is over, the program will fail to find the IMrun file and will stop. Otherwise, every half hour the program will overwrite the outfile and name the previous outfile as datafile_name.out.old.

VII. Viewing your results

To view the results, open the Excel file called 'IMchart.xls' and then open the datafile_name.out file in Excel also. In this file, look for "MARGINAL HISTOGRAMS IN DEMOGRAPHIC UNITS" and report the following values in the table below. Depending on which model you fit your data to, you may not have all of these parameters. Hopefully during the coming week you will have a change to repeat this analysis using much long run times, e.g., 12 to 48 hours, in which case you should prepare a second table and report the new values as well, and compare the two analyses!

	q1 actually N_e pop1	q2 actually N_e pop2	qA actually N_A	m1	m2	t divergence time
HiPt						
Mean						
95Lo						
95Hi						

Now check the actual marginal posterior probability distributions. Go down a few lines in the output file, and find the cell corresponding to the 0-value for q1. Select this cell and hold down the Shift key. Now go right (over) to the last column (perhaps 8 columns?) and down 999 lines to select the bottom right most cell of this block of numbers (while you are still holding the shift key). Now copy this block. Bring forward your IMchart.xls file, and select cell B4, and select Paste Special and hit OK (or maybe you can just select Paste directly?). Now select the different tabs/sheets and look at your estimated posterior probability distributions. If your spreadsheet is all messed up with the data points all falling on the zero line, one possibility is that your Excel software is expecting decimals to be indicated by commas (,) (the "normal" way), where as *IM* is giving you decimals using periods (.) (the *Gringolandia* way).

VII. Homework: a long MCMC run!

Each team should do at least one *IM* analysis run during the coming week. For this analysis you will have to change your command line somewhat. Here are some suggestions:

`-b 100000` = sets the 'burn-in' period of the MCMC to 100,000 generations.
`-fl -n 2 -g1 0.05` = Metroplis-coupled MCMC with 2 chains, the second being heated by 0.05.
`-L 6.0` = the outfile will be written every six hours.

LAB REPORT

Describe your data and describe the model you chose to analyze. Answer the following question:

(1.) Describe the posterior probability distributions for each of your parameters. Are they unimodal? Are minimum probability values at the extremes of the distribution, with values of higher probability located away from the extreme ends of the distribution? Do you think your priors were too wide (e.g, `-q1 10`), or not wide enough?
(2.) Are you happy with these results, and would you trust them? Why or why not?
(3.) Compare your two sets of results (short run vs long run). Are there big differences? If so, can you describe why?
(4.) How do your estimates of θ compare with the estimates of θ and π you obtained last week in the *dnaSP* lab?

(5.) Do you think your data conformed to the assumptions of the *IM* model and analysis? Which assumption might have been violated, and how would this violation have affected your results?

This LAB REPORT is due next Wednesday, the 29th of February.

Please also update your FINAL REPORT with the Methods and Results that came out of today's lab. In the methods sections describe the data and model (as above), as well as any assumptions (e.g., the mutation rate you assumed and how you chose that mutation rate). In the Results section, include at least one table of output (see above), preferably your second, longer analysis. Also, please begin your discussion section by writing a paragraph discussing the strengths and weaknesses of your *IM* analysis and results. Include the above two citations in your Lit Cited section, as well as any other relevant papers (e.g., on mutation rates in your organism).