database documentation: genetics

K. Mackay and P. Smith

NIWA Fisheries Data Management
database documentation series
V 1.0: 18 June
2001

Contents

1 Introduction....3
2 Genetic data collection....3
3 Data formats....4
3.1 BIOSYS-1....4
3.2 REAP....5
3.3 GENEPOP....5
4 References....6

List of figures

Figure 1: File structure of the genetics “database” 4

1 Introduction

The National Institute of Water and Atmospheric Research (NIWA) currently carries out the role of Data Manager and Custodian for the fisheries research data owned by the Ministry of Fisheries.

The Ministry of Fisheries data set incorporates historic research data, data collected more recently by MAF Fisheries prior to the split in 1995 of policy to the Ministry of Fisheries and research to NIWA, and currently data collected by NIWA and other agencies for the Ministry of Fisheries.

This document is a brief introduction to the **genetic** "database", and is part of the database documentation series produced by NIWA.

This is a "database" in the strict sense of the word, where "database" is defined as a systematic collection of computer data. However the genetics data are not stored within a Database Management System (DBMS); (e.g., EMPRESS, Oracle), but rather as a compiled collection of computer files.

This document is intended as a guide for users and administrators of the **genetic** "database".

2 Genetic data collection

Genetic data have been collected and analysed for several Ministry of Fisheries projects on orange roughy (e.g., ORH9703, DEOR13, MOF802A, FBOR01) and black and smooth oreo (e.g., DEE9801). Tissue samples are collected either fresh in the field from trawl surveys, or from whole specimens returned back to Greta Point.

The genetic materials are coded into ASCII files. The data are managed in a file system and are collated and analysed in 3 PC computer programs, with each program having their own ASCII input file format. Figure 1 shows the file system set up for the management of these genetic data. Each analysis program has its own directory, or sub-tree, for storing it's own format of input file. Within each of these directories, the data are further sub-divided by project directories, with each project directory holding all the input and output files generated for and by the relevant analysis program. Any one project may have directories in none, one or all three of the analysis program sub-trees.

Data validation is currently carried out by Dr Peter Smith (NIWA), who checks all data entries. In addition, the programs will not run, or throw up spurious alleles, if data are entered incorrectly.

3 Data formats

3.1 BIOSYS-1

BIOSYS-1 is a FORTRAN-77 program for the analysis of electrophoretically detectable allelic variation in population and biochemical systematics. The program computes allele frequencies and genetic variability measures, Hardy-Weinberg expectations, F-statistics, heterogeneity chi square analyses, and genetic distances/similarities.

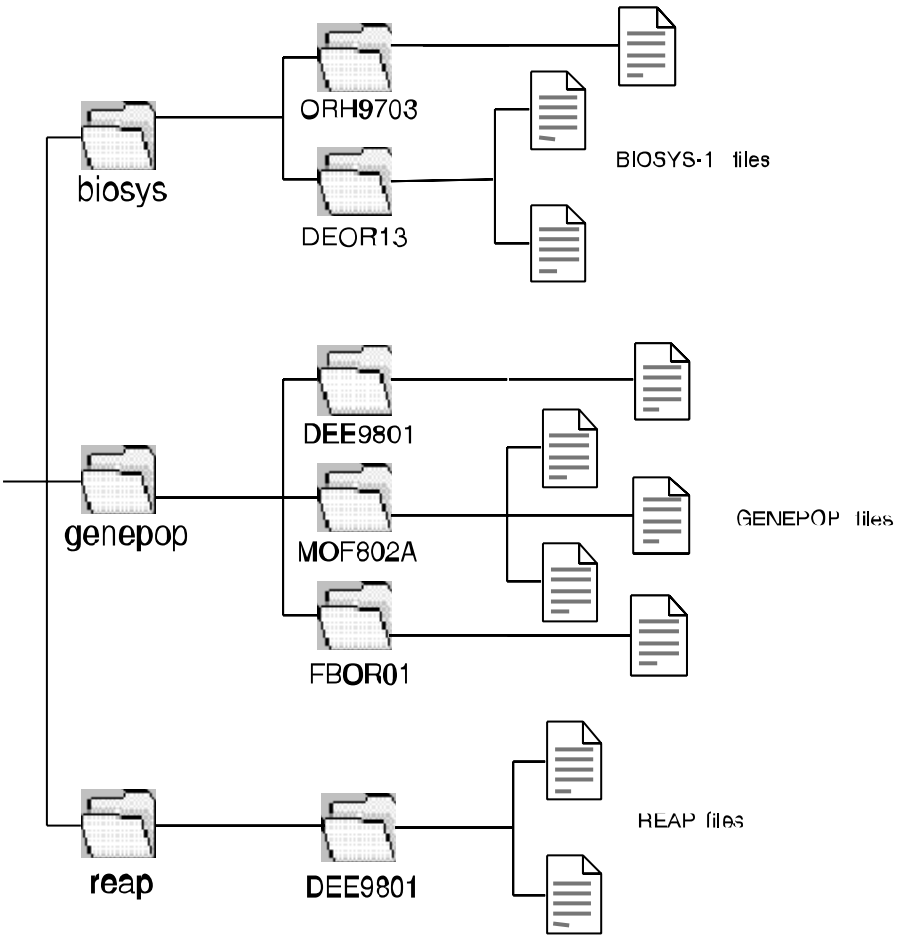

Figure 1: File structure of the genetics "database"

Inputs to BIOSYS-1 are ASCII files consisting of two basic components:

1) A brief header section, including a job title and locus labels;
2) One or more calls to "STEP" routines that give BIOSYS-1 information regarding the form of the data and the analyses that the program is to perform.

Below is a sample BIOSYS-1 input file, including header and STEP data, for a Single-Individual Genotype input:

```
SINGLE INDIVIDUAL GENOTYPE INPUT (ALPHABETIC ALLELIC DESIGNATIONS)
NOTU=1, NLOC=15,NALL=5,CRT;
(12(1X,A5)/3(1X,A5))
 LDH-1 LDH-2 MDH-1 MDH-2 IDH-1 IDH-2 GPD-1 PGM-1 PGI-1 PGI-2 SOD-1
 LAP-1 EST-1 EST-2 PEP-1
```

```
STEP DATA:
DATYP=1,ALPHA;
(A4,7X,15(1X,A1,A1))
DS1  CHATHAM RISE              1    1
0001  DS1   AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA
0002  DS1   AA AA AB AA AA AA AA AA AA AA AA AA AA AC AA
0013  DS1   AA AA AB AA AA AA AA AB AA AA AA AA AA AB AA
0014  DS1   AA AA AA AA AA AA AA AC AA AA AA AA AA AA AA
0015  DS1   AA AA AB AA AA AA AA AA AA AA AA AA AA AB AA
0016  DS1   AA AA AA AA AA AA AA AA AA AA AA AA AA AB AA
0036  DS1   AA AA AA AA AA AA AA AB AA AA AA AA AA BB AA
0037  DS1   AB AA BB AA AA AA AA AA AA AA AA AA AA AD AA
0038  DS1   AA AA AB AA AA AA AA AB AA AA AA AA AA AB AA
0039  DS1   AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA
0040  DS1   AA AA AA AA AA AA AA BB AA AA AA AA AA AC AA
NEXT
END;
```

3.2 REAP

The Restriction Enzyme Analysis Package (REAP) is a suite of nine programs (written in TURBO PASCAL 5.0 and TURBO C 1.0) designed to alleviate some of the difficulties inherent in restriction data manipulation, as well as to carry out some common phylogenetic analyses of restriction fragment or restriction site DNA data.

The user creates an ASCII file of composite haplotypes and a corresponding file of restriction enzyme profiles; from this REAP will generate a binary matrix, remove uninformative characters or Operational Taxanomic Units (OTUs), and compute estimates of evolutionary distance ($d \pm SE$) for site or fragment data. In addition, there are programs to estimate haplotype and nucleotide divergence among populations, to assess geographic heterogeneity in haplotype frequency distributions through Monte Carlo simulation, and to estimate genetic distance from DNA sequence data.

Each of the nine programs can run independently, as part of a batch process, or as a module in the integrated environment. Most of the programs can handle an unlimited number of OTUs and a maximum of 30,000 characters per OTU.

3.3 GENEPOP

GENEPOP is a population genetic software package for haploid or diploid data that is able to perform two major tasks. First, it computes exact tests or their unbiased estimation for Hardy-Weinberg equilibrium, population differentiation, and two-locus genotypic disequilibrium. Second, it converts the input GENEPOP file to formats used by other popular programs like BIOSYS, thereby allowing communication between them (ecumenicism). GENEPOP is written in QUICK BASIC and TURBO-PASCAL.

GENEPOP requires an ASCII input file. All kinds of missing data can be handled. Only two numbers code each allele, so that no more than 99 alleles can be considered. The number of populations or loci is not limiting for most options. After checking the input file, GENEPOP displays a general menu for a variety of analyses.

The main test carried out is the Hardy-Weinberg (HW) test. The HW test is performed for each locus in each population. If there are four alleles or less, the exact HW test is performed. If more than four alleles are present, an unbiased estimation of the exact HW probability is performed using the Markov chain method. In both cases, GENEPOP provides the probability of error when rejecting H_o (i.e., HW equilibrium) and, if the Markov chain method has been used, the standard error of the estimate. Other classical parameters are also automatically computed: expected genotypic proportions, allele frequencies, observed and expected numbers of homozygotes and hetrozygotes, and so on.

4 References

Raymond, M. and Rousset, F., 1995. GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. Journal of Heredity 86(3). pp. 248-249.

McElroy, D., Moran, P., Bermingham, E., and Kornfield, I., 1992. REAP: An Integrated Environment for the Manipulation and Phylogenetic Analysis of Restriction Data. Journal of Heredity 83(2). pp. 157-158.

Swofford D. and Selander, R. 1989. BIOSYS-1. A Computer Program for the Analysis of Allelic Variation in Population Genetics and Biochemical Systematics. Release 1.7. Illinois Natural History Survey Press. 43p.