A database approach to illustrate genetic trends in fishes

Maria Lourdes D. Palomares
Hydrobiologist

Christine Marie V. Casal
Hydrobiologist

Introduction

Data on finfish are available in a wide range of sources such as books, journals, proceedings, reports and computerized databases. However, these sources are often not readily available to users, be they policy-makers, scientists or students seeking information on the genetic characteristics of fishes for varied purposes. Such lack of information materials is especially felt in developing countries where meager resources do not allow for institutional or individual subscription to important journals. Thus, assembling available information into one document and making old and current data available is a significant support to scientists, researchers, policy-makers, educators and students.

Encoding data into well-structured databases is a recent approach in facilitating information gathering, dissemination and exchange. The technological advances brought about by the computer industry have improved the feasibility of working with large information systems and databases. These systems not only facilitate the dissemination of data but also provide tools for their easy manipulation and analysis. This is an important step, especially in

the field of biological science (*e.g.* in the study of biodiversity and genetics) in order to transform data into information which can then be used as a basis to educate natural resource managers and policy-makers on the status of living resources. Such information thus facilitates the definition and implementation of management measures to conserve these resources.

The question then is how to turn data into useful information. One method that can be implemented post-entry is the use of graphics in depicting relationships between any two (or more) parameters. An example of efficient and easy to implement graphical tool are scatter plots. Drawing scatter plots, however, must be preceded by a process of hypothesis building.

Some postulations involving genetic data and which were later tested using scatter plots are enumerated below. The data from the biological database known as FishBase, developed by Iclarm with support from FAO and many other collaborators and supported by the European commission, were used. The hypotheses centered on heterozygosity, polymorphism, DNA content and chromosome numbers, viz:

1. Heterozygous organisms have a higher degree of polymorphic loci than homozygous organisms and a direct (possibly linear) relationship between values of heterozygosity and polymorphism can be postulated.
2. Plotting observed *vs.* expected values of a parameter tests the predictive value of empirical formulae (SOKAL and ROHLF, 1995). Ideally, an observed value should be equal to the expected value, resulting in a scatter plot that shows a "direct" correspondence between expected and observed values, *i.e.*, a linear progression at a 45° angle from the origin of the XY axis. If a line is traced to join these points, a straight line results. This is called the 1:1 correspondence line. However, variation is intrinsic in natural processes, hence observed actual values deviate from the 1:1 line, but should remain close to it. Outliers merit scrutiny.
3. HINEGARDNER and ROSEN (1972) presented haploid DNA content data for almost 300 teleost species and showed that more specialized (or evolutionally advanced) fishes have less DNA than more generalized forms. This trend was further verified by CUI *et*

al., (1991) who determined the cellular DNA content of 42 species of Chinese freshwater fishes.
4. Since chromosomes carry the genetic material, *i.e.* DNA, then it can be postulated that the number of chromosomes is directly related to the DNA content.

Materials and methods

Three types of genetic data (continuous numeric variables) were investigated here, through FishBase (FROESE and PAULY 1996). The first study used heterozygosity and polymorphism values derived from allele frequency data obtained from the literature on electrophoretic studies for 195 species (*i.e.* 10,900 records), extracted from about 40 references. The other study compared the DNA content of over 350 species and the chromosome numbers of over 1300 species, extracted from over 400 references.

Heterozygosity and Polymorphism

Two graphs were created: one to plot heterozygosity against polymorphism and the other to plot observed heterozygosity against expected heterozygosity. Data points for a chosen species were overlaid against all other species for which data existed.

Heterozygosity is defined as the proportion of heterozygotes for a given locus in a population. Heterozygous individuals are diploid organisms that have inherited different alleles from each parent, *i.e.* they carry different alleles at the corresponding places on paired chromosomes. Heterozygosity for each locus was computed from allele frequency data using the Hardy-Weinberg equation (CARPENA *et al.*, 1993) expressed as: $p^2+2pq+q^2=1$.

Where p is the frequency of the dominant allele and where q is the frequency of the recessive allele. The expected heterozygosity was

computed by averaging values of heterozygosity per locus over all loci studied for a particular population.

A polymorphic locus hosts two or more different alleles (LAWRENCE 1995). The percentage of polymorphic loci was calculated as:

% polymorphism = number of polymorphic loci/total number of loci studied * 100

Figure 1 was created to test this hypothesis for finfish. Note that the plot was made using expected and not observed heterozygosity, because not all of the sources used in FishBase provided estimates for this variable.

Expected and observed heterozygosity values were plotted following these above assumptions using only populations for which both values were available. Data points which deviated from the 1:1 line, *i.e.* beyond an imaginary 95% confidence interval, were verified. The original data source was inspected for possible discrepancies arising from encoding errors.

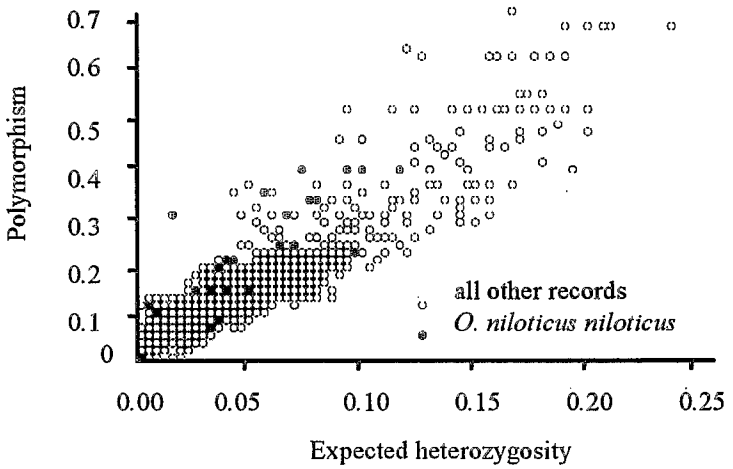

Figure 1
Scatter plot of polymorphism *vs.* expected heterozygosity for *Oreochromis niloticus niloticus* printed from FishBase (Froese and Pauly 1996). Note outlier standing out from the group of black dots with polymorphism = 0.3.

DNA content, chromosome number and phylogenetic order

To test this hypothesis, a scatter plot was created for the DNA content (expressed as the haploid value, in picograms) against the phylogenetic order of families presented by NELSON (1994). The DNA content data was obtained from the Genetics table of FishBase. Assuming that this relationship is also true for chromosome numbers, diploid chromosome numbers were plotted against the phylogenetic order of families. The chromosome data was also obtained from the Genetics table of FishBase.

Results and discussion

Heterozygosity and Polymorphism

The scatter plots for expected heterozygosity vs. polymorphism are presented in Figure 1 for *Oreochromis niloticus niloticus* (Linneaus, 1758). A strong increasing trend is evident with data points concentrated in the center of the graph and around an imaginary 1:1 correspondence line. Note that one (in parenthesis) of the 27 data points for *O. n. niloticus* deviates from the general trend. Verification of the record entered against the original source confirmed that an error was made in encoding the allele frequency of one locus. Scatter plots for observed *vs.* expected heterozygosity are presented in Figure 2. It is evident that one particular population of *O. n. niloticus* (in parenthesis) showed a very large difference between expected and observed heterozygosity. Verification of the record entered showed that the necessary linkage between allele frequencies and the publication used for some of the loci recorded for this species in this specific study were erroneous, *i.e.* the allele frequency records were linked to the wrong population. This resulted in erroneous computations of expected heterozygosity.

These results confirm, as expected, a direct relationship between heterozygosity and polymorphism and showed the importance of investigating outliers.

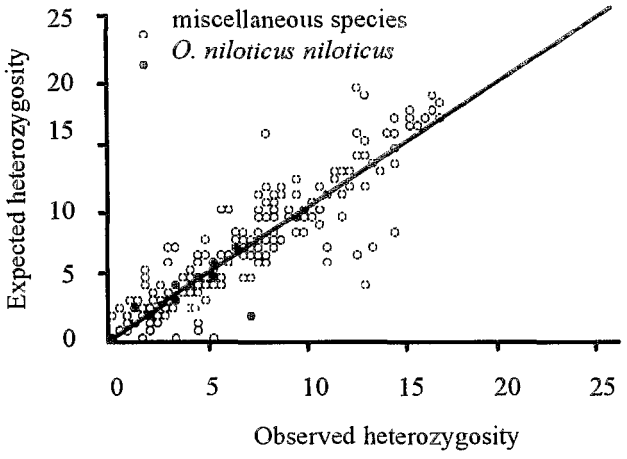

Figure 2
Scatter plot of expected vs. observed heterozygosity for *Oreochromis niloticus niloticus* printed from FishBase (Froese and Pauly 1996). Note outlier standing out from the group of black dots close to the 1:1 correspondence line.

DNA content, chromosome number and phylogenetic order

Scatter plots of the phylogenetic order of family against DNA contents and chromosome numbers are presented in Figures 3 and 4. These figures show two apparent trends: (1) a decrease in DNA content/chromosome number in the course of evolution; and (2) a decrease in variability of DNA content/chromosome numbers of species within orders of increasing phylogenetic order. Some very high values of DNA content (4.0-4.5) and chromosome number (250-300) identified in these graphs belong to polyploid groups

(*e.g.* Gasterosteidae). The following studies support the observed trends, viz.:

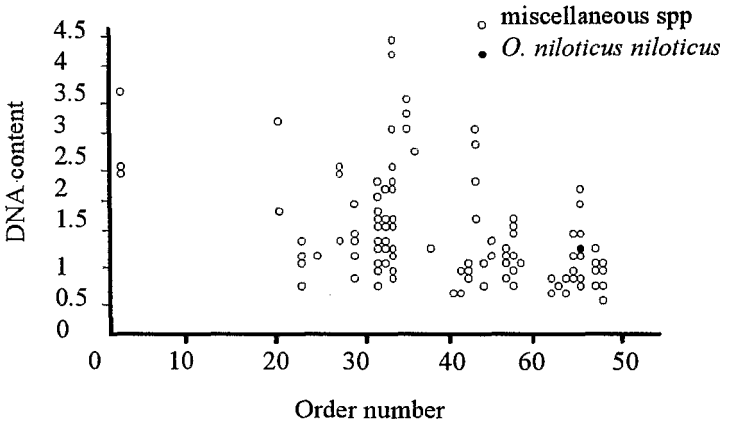

Figure 3
Scatter plot of DNA content *vs* phylogenetic order printed from FishBase (FROESE and PAULY 1996).

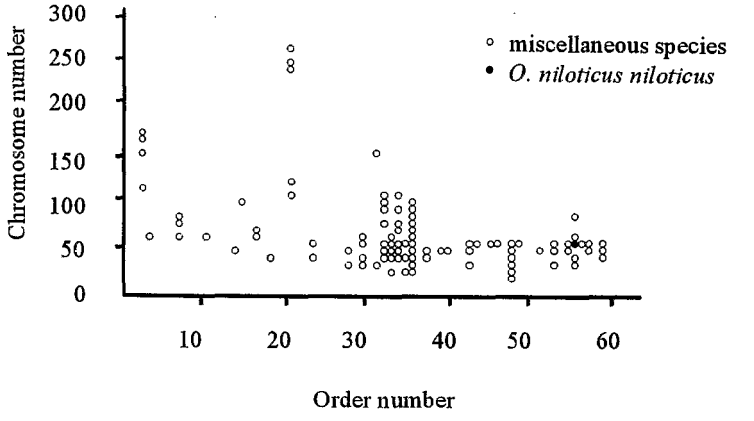

Figure 4
Scatter plot of chromosome number *vs* phylogenetic order printed from FishBase (FROESE and PAULY 1996).

- karyotypes with large numbers of chromosomes (HINEGARDNER and ROSEN, 1972) and including a large proportion of telocentric chromosomes are more representative of a primitive elasmobranch genome than are other karyotypes (SCHWARTZ and MADDOCK, 1985);
- selachians have high chromosome numbers (2n=50-100) which decrease in more specialized species through a loss of acrocentrics and microchromosomes (STINGO and CAPRIGLIONE, 1985);
- within a taxonomic group, an increase followed by a gradual decrease in DNA, which is associated with specialization, appears to have accompanied fish evolution (HINEGARDNER and ROSEN, 1972).

Data quality control

As was shown above, the task of compiling biological databases is large and complex. The quality of the information being entered (and therefore to be disseminated) must be assured. The most common method is to ask collaborators or experts to verify the data entered. This is a time consuming work, because the amount of information to be verified can comprise tens of thousands of records. It is thus difficult to ask "volunteer" collaborators to put their own work aside and to spend days or weeks verifying rows and rows of encoded data.

The outliers, *i.e.*, data points outside the general (expected) pattern of a relationship identified here resulted from (a) an encoding (human) error (Figure 1); (b) a source code level (programming) error (Figure 2). Errors in the original data source might also occur, but no example could be identified with the data set used. Note that only very rarely would outliers indicate a "way-out" fish species because fish are all "built" according to a common design pattern, dictated by the laws of thermodynamics, etc. Scatter plots permit the efficient verification of large volumes of data in a short period of time (usually a few minutes depending on computer speed) and provide a picture of how the encoded data, taken together, behave.

Conclusions

Genetic data compiled to date in FishBase allowed to test and verify the relationship between heterozygosity and polymorphism. The equation to estimate expected heterozygosity was confirmed in a 1:1 plot over observed heterozygosity. A plot of DNA content and chromosome number of the phylogenetic rank of fish orders showed a decreasing trend and confirmed some predictions from recent literature. Furthermore, scatter plots turned out to be also useful tools in identifying errors, which can thus be verified and repaired. Such functions, if habitually incorporated in databases, permit the automatic and regular verification of data being encoded, thus improving the quality of the data stored.

References

CARPENA (A.L.), ESPINO (R.R.C.), ROSARIO (T.L.), LAUDE (R.P.), 1993 — *Genetics at the population level.* SEARCA, UPLB, Laguna, Philippines, 244 p.

CUI (J. X.), YU (R.Q.), 1991 — Nuclear DNA content variation in fishes. *Cytologia*, 56(3): 425-429.

FAN (Z.), FOX (D.P.), 1991 — Robertsonian polymorphism in plaice, *Pleuronectes platessa* L. and cod, *Gadus morhua* L. (Pisces, Pleuronectiformes and Gadiformes). *J. Fish Biol.*, 38(5): 635-640.

FROESE (R.), PAULY (D.), 1996 — (eds). *FishBase 96. Concepts, design and data sources.* Iclarm, Manila, Philippines, 179 p.

HINEGARDNER, (R.) and (D.E.) ROSEN. 1972 — Cellular DNA content and the evolution of teleostean fishes. *American Naturalist*, 106(951):621-644.

KLINKHART (M. M.), TESCHE (H.) GREVEN, 1995 — (eds.) *Database of fish chromosomes.* Westarp Wissenschaften, 237 p.

LAWRENCE (E.), 1995 — (ed.) *Henderson's dictionary of biological terms.* 11th Edition. Longman Singapore Publishers (Pte) Ltd., Singapore. 693 p.

SOKAL (R.R.), ROHLF (F.J.), 1995 — *Biometry. The principles and practice of statistics in biological research.* Third Edition. W.H. Freeman and Company, New York, 887 p.

STINGO (V.), CAPRIGLIONE (T.), 1985 — DNA and chromosomal evolution in cartilaginous fish. In UYENO (T.), ARAI (R.), TANIGUCHI (T.), MATSUURA (K.), (eds.) *Indo-Pacific Fish Biology.*

Proceedings of the second international conference on Indo Pacific Fishes, Tokyo National Museum, Ueno Park, Tokyo, July 29-Aug. 3, 1985. p140-147

SCHWARTZ (F.J.), MADDOCK (M.B.), 1985 — Comparisons of karyotypes and cellular DNA contents within and between major lines of elasmobranchs. In UYENO (T.), ARAI (R.), TANIGUCHI (T.), MATSUURA (K.), (eds.) *Indo-Pacific Fish Biology*. *Proceedings of the second international conference on Indo Pacific Fishes*, Tokyo National Museum, Ueno Park, Tokyo, July 29-Aug. 3, 1985. p148-157

WARD (R.D.), WOODWARK (M.), SKIBINSKI (D.O.F.), 1994 — Comparison of genetic diversity levels in marine, freshwater and anadromous fishes. *J. Fish Biol.*, 44: 213-232.