

10. ——— and B. G. SANDERS. Developmental profile of chicken splenic lymphocyte responsiveness to Con A and PHA and studies on chicken splenic and bone marrow cells capable of inhibiting mitogen stimulated blastogenic responses of adult splenic lymphocytes. *J. Immunol.* 125:1792-1797. 1980.
11. MIGGIANO, V., M. NORTH, A. BUDER, and J. R. L. PINK. Genetic control of the response of chicken leukocytes to a T-cell mitogen. *Nature* 263:61-63. 1976.
12. ROSE, N. R., L. D. BACON, R. S. SUNDICK, and W. E. BRILES. The significance of cellular alloantigens in spontaneous autoimmune thyroiditis (SAT) of OS chickens. In *Animal Models of Comparative and Developmental Aspects of Immunity and Disease*. M. E. Gershwin and E. L. Cooper, Eds. Pergamon Press, NY. p. 143-153. 1978.
13. SANDERS, B. G. and K. KLINE. IgG immunoglobulin deficiency in muscular dystrophic chickens. *J. Hered.* 68:55-56. 1977.
14. ——— and ———. Immune involvement in myopathies: emphasis on the inherited muscular dystrophy chicken model. In *Animal Models of Comparative and Developmental Aspects of Immunity and Disease*. M. E. Gershwin and E. L. Cooper, Eds. Pergamon Press, NY. p. 1221-1231. 1978.
15. ———, ———, and C. J. MORTON. Serum IgG levels in hereditary muscular dystrophic chickens. *Biochem. Genet.* in press. 1981.

tionships within and among natural populations and species. A major advantage of electrophoretic approaches is that the same data base can be applied to a wide variety of problems. Consequently, researchers are heavily dependent on high-speed computers for the analysis of these data.

Unfortunately, despite the broad applicability of allozymic information, most currently available computer programs are somewhat restricted in scope. Typically, a researcher uses one program to compute genetic variability parameters, one or more other programs to calculate various indices of genetic similarity and/or distance, and still other programs to generate dendrograms via cluster analysis or other tree-building procedures. Aside from the obvious inefficiency of this approach, the use of several programs also tends to promote inaccuracy due to errors in the transcription of data.

BIOSYS-1 is a single, multi-purpose program that performs most types of data analysis commonly employed by biochemical population geneticists and systematists, and thus alleviates most of the above problems.

Capabilities

Any or all of the following analyses may be performed in a single run of the program.

Computation of allele (electromorph) frequencies

The program accepts data in either of two forms for the computation of allele frequencies: 1) data are entered for each individual in the study, giving the individual's genotype at each locus for which it was scored; or 2) data are entered for each locus in each population, specifying all genotypes observed at the locus and their corresponding frequencies. Alternatively, if allele frequencies are already known, they can be entered directly for use in subsequent program steps.

Measures of genetic variability

The following measures of genetic variability (and their standard errors) are computed for each population sample analyzed: mean number of alleles per locus, percentage of loci polymorphic using 95 percent, 99 percent, and no criteria for polymorphism (frequency of most common allele less than or equal to 0.95, 0.99, or 1.00), and mean heterozygosity (both direct-count and Hardy-Weinberg expected, including Nei's⁹ unbiased estimate). Allele frequencies and variability measures may be printed either population by population (all variability measures; see Figure 1A) or in single tables (user-selected variability measures; see Figure 1B and C).

Hardy-Weinberg equilibrium

Each polymorphic locus is tested for conformance of genotype frequencies to those expected under Hardy-Weinberg equilibri-

um. Expected frequencies are calculated using Levene's⁶ correction for small sample size. A chi-square goodness-of-fit test is performed and the level of significance (*P* value) determined. If more than two alleles are present in a population sample, certain genotypic classes are pooled (see Figure 1D) and the test performed again. In addition the genotypic fixation index¹⁷ (*F*) and Selander's¹³ *D* statistic for the excess or deficiency of heterozygotes are calculated for each polymorphic locus.

F statistics

Wright's^{17,18} *F* statistics for the analysis of population structure by standardized genetic variance are computed for all variable loci. Up to four hierarchical levels are permitted. In general, the subdivisions of Wright¹⁸ are used: "demes" (*D*), "regions" (*R*), "subdivisions" (*S*), and "total range" (*T*). This permits calculation of the following statistics: *F_{DR}*, *F_{RS}*, *F_{ST}*, *F_{DS}*, *F_{RT}*, and *F_{DT}*. However, any combination of hierarchical levels is permitted (e.g., populations, subspecies, species, and genus or subpopulations, local races, and total range).

Heterogeneity chi-square

The significance of interpopulational heterogeneity in allele frequencies is evaluated using a heterogeneity chi-square test (see Workman and Niswander¹⁰). The entire array of populations may be analyzed or subsets of this array may be analyzed independently.

Similarity and distance coefficients

The following indices are computed: Rogers¹² genetic similarity (*S*) and distance (*D*), Nei⁸ standard genetic identity (*I*) and distance (*D*), Nei's¹⁰ minimum distance (*D_m*) and Nei's more recent "unbiased" versions of *I*, *D*, and *D_m*. In addition, the less commonly used, although highly appropriate, distances of Prevosti (see Wright¹⁶) and Cavalli-Sforza and Edwards³ are calculated. Values are printed as matrices representing all possible pairwise comparisons between populations. The arrangement of these matrices (e.g., Rogers *S* above diagonal, Nei *D* below diagonal; see Figure 1E) may be specified by the user. Any number of matrix pairs may be printed.

In cases where several populations per species are sampled, averages for intraspecific and interspecific comparisons can be calculated (Figure 1F). In addition, the possible range (0 to 1) of similarity coefficients is divided into 20 equal intervals and the relative frequency distribution of single-locus coefficients on this scale is computed. This is done separately for comparisons involving: 1) conspecific populations; 2) congeneric species; and 3) different genera (see Avise and Smith¹ for applications).

Cluster analysis

Cluster analysis may be performed on any of the above matrices using either the

The Journal of Heredity 72:281-283. 1981.

BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics

David L. Swofford and
Richard B. Selander

ABSTRACT: BIOSYS-1 is a FORTRAN IV program designed to aid biochemical population geneticists and systematists in the analysis of electrophoretically detectable allelic variation. It can be used to compute allele frequencies and genetic variability measures, to test for deviation of genotype frequencies from Hardy-Weinberg expectations, to calculate *F*-statistics, to perform heterogeneity chi-square analysis, to calculate a variety of similarity and distance coefficients, and to construct dendrograms using cluster analysis and Wagner procedures. The program, documentation, and test data are available from the authors.

RECENT YEARS have witnessed an explosion in the use of electrophoretic techniques to assess levels of variability and genetic rela-

The authors are, respectively, graduate student and professor in the Department of Genetics and Development at the University of Illinois, Urbana, IL 61801. They wish to thank Gregory S. Whitt for helpful comments on the manuscript. Computer time for development of this program was provided by the University of Illinois Research Board.
© 1981, American Genetic Association.

ALLELE FREQUENCIES AND GENETIC VARIABILITY MEASURES

POPULATION: BIG RIVER, IL (DB1)

ALLELE	LOCUS AND SAMPLE SIZE														
	LDH-1 11	LDH-2 11	MDH-1 11	MDH-2 11	IDH-1 11	IDH-2 11	OPD-1 11	PGH-1 11	POI-1 11	POI-2 11	SOD-1 11	LAP-1 11	EST-1 11	EST-2 11	PEP-1 11
A	.955	1.000	.727	1.000	1.000	1.000	1.000	.727	1.000	1.000	1.000	1.000	1.000	.591	1.000
B	.045	0.000	.273	0.000	0.000	0.000	0.000	.273	0.000	0.000	0.000	0.000	0.000	.273	0.000
C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	.045	0.000	0.000	0.000	0.000	0.000	.091	0.000
D	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	.045	0.000
H	.087	0.000	.397	0.000	0.000	0.000	0.000	.417	0.000	0.000	0.000	0.000	0.000	.566	0.000
H(UMB)	.091	0.000	.414	0.000	0.000	0.000	0.000	.437	0.000	0.000	0.000	0.000	0.000	.593	0.000
H(D.C.)	.091	0.000	.364	0.000	0.000	0.000	0.000	.364	0.000	0.000	0.000	0.000	0.000	.636	0.000

MEAN HETEROZYGOUSITY PER LOCUS (BIASED ESTIMATE) = .098 (S.E. .050)
 MEAN HETEROZYGOUSITY PER LOCUS (UNBIASED ESTIMATE) = .102 (S.E. .052)
 MEAN HETEROZYGOUSITY PER LOCUS (DIRECT COUNT ESTIMATE) = .097 (S.E. .051)
 MEAN NUMBER OF ALLELES PER LOCUS = 1.47 (S.E. .24)
 PERCENTAGE OF LOCI POLYMORPHIC (0.95 CRITERION) = 20.0
 PERCENTAGE OF LOCI POLYMORPHIC (0.99 CRITERION) = 26.7
 PERCENTAGE OF LOCI POLYMORPHIC (NO CRITERION) = 26.7

GENETIC VARIABILITY AT 15 LOCI IN ALL POPULATIONS

(STANDARD ERRORS IN PARENTHESES)

POPULATION	MEAN SAMPLE SIZE PER LOCUS	MEAN NO. OF ALLELES PER LOCUS	PERCENTAGE OF LOCI POLYMORPHIC	MEAN HETEROZYGOUSITY	
				DIRECT COUNT	HWBBD EXPECTED
1. BIG RIVER, IL (.0.0)	11.0 (.0.0)	1.5 (.1.2)	20.0	.097 (.051)	.102 (.052)
2. LITTLE CREEK, IN (.0.0)	6.0 (.0.0)	1.2 (.1.2)	20.0	.078 (.043)	.068 (.037)
3. SMALL BRANCH, NY (.0.0)	4.0 (.0.0)	1.0 (.0.0)	0.0	0.000 (0.000)	0.000 (0.000)
4. MTN. LAKE, CO (.0.0)	6.0 (.0.0)	1.3 (.1.3)	26.7	.100 (.053)	.108 (.052)
5. CLEAR POND, CO (.3)	3.3 (.3)	1.1 (.1)	13.3	.050 (.036)	.045 (.032)
6. MUDDY BROOK, OH (.1)	6.9 (.1)	1.1 (.1)	13.3	.040 (.030)	.044 (.034)

A LOCUS IS CONSIDERED POLYMORPHIC IF THE FREQUENCY OF THE MOST COMMON ALLELE DOES NOT EXCEED .95

UNBIASED ESTIMATE (SEE NEI, 1978)

CHI-SQUARE TEST WITH POOLING

POPULATION: BIG RIVER, IL (DB1)

LOCUS	CLASS	OBSERVED FREQUENCY	EXPECTED FREQUENCY	CHI-SQUARE	DF	P
POI-1	HOMOZYGOUS FOR MOST COMMON ALLELE	6	5.714	.200	1	.655
	COMMON/RARE HETEROZYGOUS	4	4.571			
	RARE HOMOZYGOUS AND OTHER HETEROZYGOUS	1	.714			
EST-2	HOMOZYGOUS FOR MOST COMMON ALLELE	3	3.714	.801	1	.371
	COMMON/RARE HETEROZYGOUS	7	5.571			
	RARE HOMOZYGOUS AND OTHER HETEROZYGOUS	1	1.714			

MATRIX OF GENETIC SIMILARITY AND/OR DISTANCE COEFFICIENTS

BELOW DIAGONAL: ROBERS (1972) GENETIC DISTANCE
 ABOVE DIAGONAL: NEI (1978) UNBIASED GENETIC DISTANCE

POPULATION	1	2	3	4	5	6
1 BIG RIVER, IL	#####	.048	.015	.298	.381	.798
2 LITTLE CREEK, IN	.095	#####	.040	.447	.520	.717
3 SMALL BRANCH, NY	.062	.072	#####	.417	.488	.894
4 MTN. LAKE, CO	.105	.398	.364	#####	.085	.611
5 CLEAR POND, CO	.140	.423	.390	.121	#####	.554
6 MUDDY BROOK, OH	.551	.523	.589	.471	.434	#####

MATRIX OF SIMILARITY/DISTANCE COEFFICIENTS AVERAGED BY SPECIES

COEFFICIENT: ROBERS (1972) GENETIC SIMILARITY

SPECIES	NO. OF POP.	1	2	3
1 SMITHI	3	.924 (.905-.938)		
2 JONESI	2	.630 (.577-.693)	.879 (.879-.879)	
3 DOEII	1	.446 (.411-.477)	.548 (.529-.566)	##### (#####)

ONLY ONE POPULATION INCLUDED WITHIN SPECIES AVERAGE WITH ITSELF

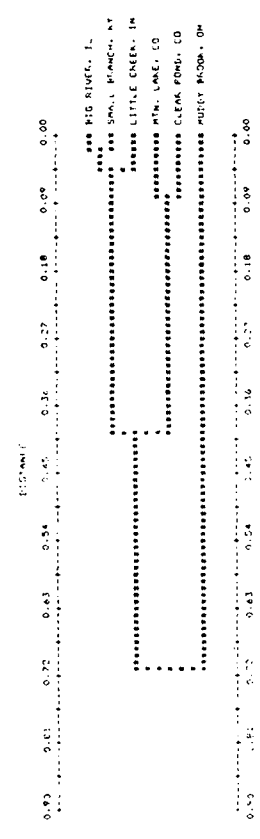
FIGURE 1 Sample table output from BIOSYS-1. A—allele frequencies and variability measures for a single population. B—allele frequencies for all populations. C—selected variability measures for all populations.

D—Hardy-Weinberg test with pooling. E—genetic similarity/distance matrix. F—genetic distance coefficients averaged by species (conspecific comparisons along the diagonal).

B

BIOSYS-1: LINE FROM UNROOTED PAIR, PAIR, SELECTED ANNOTATES

 COMPUTED USER=NEI 1978B UNROOTED GENETIC DISTANCE (D)

**A**

DISTANCE WAGNER USING ROGERS DISTANCES

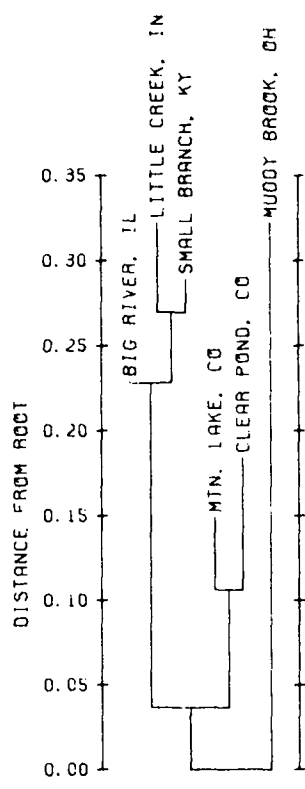


FIGURE 2 Sample tree output from BIOSYS-1. A—distance Wagner tree on Calcomp plotter (branch lengths proportional to inferred patristic distances). B—UPGMA phenogram on line printer.

weighted or unweighted pair group methods with recomputation of coefficients by arithmetic averaging (WPGMA and UPGMA methods, respectively)¹⁴. The resulting dendrograms are drawn on the Calcomp plotter or printed on the line printer (Figure 2A and B, respectively). Goodness-of-fit of the dendrogram to the input matrix is evaluated using the cophenetic correlation coefficient¹⁴ or the F value of Prager and Wilson¹¹.

Distance Wagner procedure

The method of Farris⁵ is used to construct Wagner networks from any of the above distance matrices. The networks may be rooted by either of two methods⁵: 1) the root is placed at the midpoint of the greatest patristic difference separating a pair of populations; or 2) the root is specified by the point where one or more user-designated "outgroup" populations join the network. The resulting dendrograms are output on the line printer and/or Calcomp plotter. Goodness-of-fit is evaluated as in cluster analysis (above).

If desired, a modification of the Farris procedure may be used¹⁵. Trees constructed by this modified procedure, which uses multiple addition criteria and branch length optimization, are generally superior in goodness-of-fit to those generated by the original Farris procedure.

Conventional Wagner procedure

The "rootless" or "simple" algorithms of Farris⁴ are used to construct Wagner networks from character-state data. Either binary coding⁷ or frequency coding² of allozyme data may be specified. Networks are optimized using a newly developed procedure and rooted as in the distance Wagner procedure. Trees are printed and/or drawn as above.

Discussion

A valuable feature of the BIOSYS-1 program is its ability to accept single-individual genotype data as input. Data entry can therefore begin early and can be continued throughout the progress of the electrophoretic analyses. Since the program computes allele frequencies each time it is run, the researcher can obtain a complete analysis of the available data at any point in the study. The necessity of manually recomputing allele frequencies and preparing them for program input when new data are obtained is thus eliminated. The use of variable formatting permits skipping data for loci scored initially but subsequently dropped from the study. Despite the obvious advantages of this system, the program is flexible in that data can be entered as allele frequencies if the user so desires.

BIOSYS-1 is currently dimensioned for maxima of 40 populations, 30 loci, and 10 alleles per locus, and as such requires about 300K bytes of central memory. These upper

limits can be easily decreased to lower memory requirements or increased to accommodate more data. Instructions for redimensioning are provided. The program is written in IBM FORTRAN IV and is compatible with either the FORTRAN G or WATFIV compilers. It is currently running on the IBM 360/75 and 4341 computers at the University of Illinois. A CDC FORTRAN version is also available.

A program listing, documentation, and test data may be obtained from the authors. The program is also available on punched cards or magnetic tape. Separate programs for cluster analysis and Wagner tree construction from user-supplied matrices are also available on request.

References

1. AVISE, J. C. and M. H. SMITH. Gene frequency comparisons between sunfish (*Centrarchidae*) populations at various stages of evolutionary divergence. *Syst. Zool.* 26:319-335. 1977.
2. BUTH, D. G. Biochemical systematics of the cyprinid genus *Notropis*. I. The subgenus *Luxilus*. *Biochem. Syst. Ecol.* 7:69-79. 1979.
3. CAVALLI-SFORZA, L. L. and A. W. F. EDWARDS. Phylogenetic analysis: models and estimation procedures. *Evolution* 21:550-570. 1967.
4. FARRIS, J. S. Methods for computing Wagner trees. *Syst. Zool.* 19:83-92. 1970.
5. ———. Estimating phylogenetic trees from distance matrices. *Am. Nat.* 106:645-668. 1972.
6. LEVENE, H. On a matching problem arising in genetics. *Ann. Math. Stat.* 20:91-94. 1949.
7. MICKEVICH, M. F. and M. S. JOHNSON. Congruence between morphological and allozyme data in evolutionary inference and character evolution. *Syst. Zool.* 25:260-270. 1976.
8. NEI, M. Genetic distance between populations. *Am. Nat.* 106:283-292. 1972.
9. ———. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583-590. 1978.
10. ——— and A. K. ROYCHOUDHURY. Sampling variances of heterozygosity and genetic distance. *Genetics* 76:379-390. 1974.
11. PRAGER, E. M. and A. C. WILSON. Construction of phylogenetic trees for proteins and nucleic acids: comparison of alternative matrix methods. *J. Mol. Evol.* 11:129-142. 1978.
12. ROGERS, J. S. Measures of genetic similarity and genetic distance. *Studies in Genetics*. Univ. Texas Publ. 7213:145-153. 1972.
13. SELANDER, R. K. Behavior and genetic variation in natural populations. *Am. Zool.* 10:53-66. 1970.
14. SNEATH, P. H. A. and R. R. SOKAL. Numerical Taxonomy. W. H. Freeman, San Francisco. 1973.
15. SWOFFORD, D. L. On the utility of the distance Wagner procedure. In *Advances in Cladistics: Proceedings of the First Meeting of the Willi Hennig Society*. V. A. Funk and D. R. Brooks, Eds., Publ. N.Y. Botanical Garden, NY, in press. 1981.
16. WORKMAN, P. L. and J. D. NISWANDER. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. *Am. J. Human Genet.* 22:24-49. 1970.
17. WRIGHT, S. *Evolution and the Genetics of Populations*. Vol. 2, The Theory of Gene Frequencies. University of Chicago Press, Chicago. 1969.
18. ———. *Evolution and the Genetics of Populations*. Vol. 4, Variability Within and Among Natural Populations. University of Chicago Press, Chicago. 1978.