



Hunt Institute for Botanical Documentation
5th Floor, Hunt Library
Carnegie Mellon University
4909 Frew Street
Pittsburgh, PA 15213-3890
Telephone: 412-268-2434
Email: huntinst@andrew.cmu.edu
Web site: www.huntbotanical.org

The Hunt Institute is committed to making its collections accessible for research. We are pleased to offer this digitized item.

Usage guidelines

We have provided this low-resolution, digitized version for research purposes. To inquire about publishing any images from this item, please contact the Institute.

Statement on harmful and offensive content

The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences. Due to the wide range of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions. The Hunt Institute for Botanical Documentation does not endorse the views expressed in these materials, which are inconsistent with our dedication to creating an inclusive, accessible and anti-discriminatory research environment. Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections' creators.

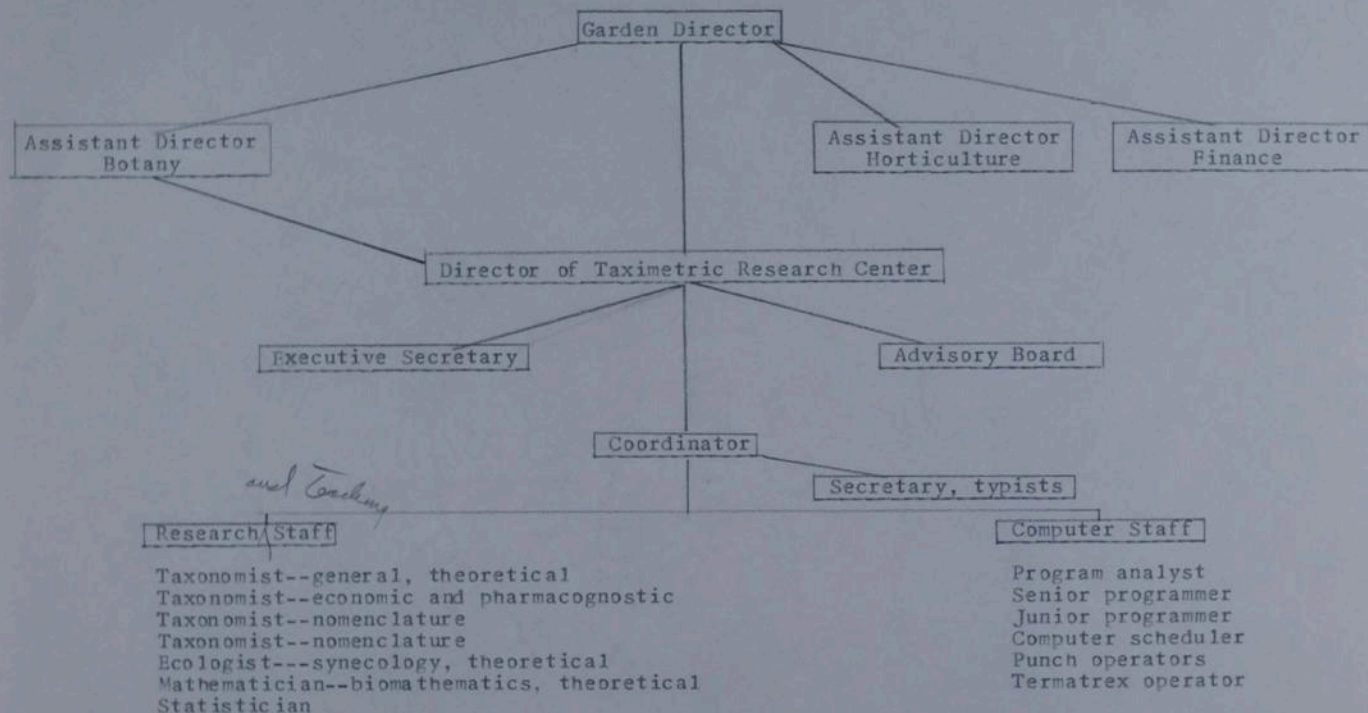
Many of the historical collections in the Hunt Institute Archives contain personal correspondence, notes, recollections and opinions, which may contain language, ideas or stereotypes that are offensive or harmful to others. These collections are maintained as records of the individuals involved and do not reflect the views or values of the Hunt Institute for Botanical Documentation or those of Carnegie Mellon University.

About the Institute

The Hunt Institute for Botanical Documentation, a research division of Carnegie Mellon University, specializes in the history of botany and all aspects of plant science and serves the international scientific community through research and documentation. To this end, the Institute acquires and maintains authoritative collections of books, plant images, manuscripts, portraits and data files, and provides publications and other modes of information service. The Institute meets the reference needs of botanists, biologists, historians, conservationists, librarians, bibliographers and the public at large, especially those concerned with any aspect of the North American flora.

Hunt Institute was dedicated in 1961 as the Rachel McMasters Miller Hunt Botanical Library, an international center for bibliographical research and service in the interests of botany and horticulture, as well as a center for the study of all aspects of the history of the plant sciences. By 1971 the Library's activities had so diversified that the name was changed to Hunt Institute for Botanical Documentation. Growth in collections and research projects led to the establishment of four programmatic departments: Archives, Art, Bibliography and the Library.

Taximetric Research Center
Personnel Chart--Organization



TAXIMETRICS RESEARCH PROGRAM

TAXIMETRICS SECTION

1. Mathematical and statistical models for biology
2. Development of computer methods for the simulation of biological concepts.
3. Classification of organisms.

TRAINING SECTION

1. Biological disciplines
2. Mathematics and applications
3. Programming

INFORMATION RETRIEVAL

1. Nomenclature of plants and animals
2. Classification of useful plant literature
3. Drug and plant correlation

GARDEN SERVICES

1. Plant records
2. Library records
3. Herbarium records
4. Business and accounting

OUTSIDE SERVICES

1. Collaborators
2. Research associates
3. Information retrieval centers

Biology and Mathematical Problems
Outline

1. Historical
 - A. Attempts at numerical handling of non-numerical information not new.
 - B. Efforts of Sewell Wright, R. A. Fisher, Gregg, et al.
 - C. Development of computers stimulated investigations of math. models that have applicability in biological problems.
 - D. Interest for taxonomy--three nearly simultaneous investigations: Sneath on bacteria; Sokal & Michener on insects, and ours for higher plants. All with programs for computers.
2. Development of our program
 - A. Studies of numerous variants of *Manihot esculenta*.
 - B. Interest of Tanimoto--how we began. Conversations, similarity of interests.
 - C. Apparent that classification is general value, but models can best be tested where finite samples available to check the model--the herbarium (museum) specimen.
 - D. The classification process--how it works. Refer to diagram.
3. Applications in many fields: archeological to zoological. Many investigations appropriate in these areas.

Questions for Seminar

1. What problems arise in biology that require or can be more effectively treated through mathematical investigation?
2. Why should a botanical garden take up the types of research that involve mathematics?
3. What sort of a group of investigators would be needed to make a sound investigation of biological-mathematical concepts?
4. Could a computer facility be justified in a botanical garden with museum and laboratory research type functions?
5. How could a new endeavor with sufficient staff and required hardware be financed at such an institution?
6. What sort of collaborative endeavors could be arranged between this center and sister institutions and endeavors?

AGENDA
SEMINAR ON MATHEMATICS AND BIOLOGY
THE NEW YORK BOTANICAL GARDEN
TUESDAY, JUNE 2, 1964

10:15 A.M. Welcome and Introduction to The New
York Botanical Garden.

Dr. William C. Steere

10:45 A.M. Introduction to Problems in Biology
and Mathematics.

Dr. David J. Rogers

11:15 A.M. The Taxonomic Facilities of The New
York Botanical Garden

Dr. Bassett Maguire

11:30 A.M. The Taximetric Model: its Implica-
tions, its Flaws, its Future Modifi-
cations.

Mr. Henry Fleming

12:00 Noon Luncheon

1:30 P.M. Mathematician's View of Biological
Problems.

Dr. Taffee T. Tanimoto

2:00 P.M. Discussion of Listed Questions

Dr. Steere presiding

4:00 P.M. Summary and second thoughts

4:30 P.M. Close

Questions for Seminar

1. What problems arise in biology that require or can be more effectively treated through mathematical investigation?
2. Why should a botanical garden take up the types of research that involve mathematics?
3. What sort of a group of investigators would be needed to make a sound investigation of biological-mathematical concepts?
4. Could a computer facility be justified in a botanical garden with museum and laboratory research type functions?
5. How could a new endeavor with sufficient staff and required hardware be financed at such an institution?
6. What sort of collaborative endeavors could be arranged between this center and sister institutions and endeavors?

WG 5-3000
Kitchin Rd.

H. H. Goldstine
Thomas J. Watson Research Center
Yorktown Heights, New York

June 16, 1964

Dr. David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx Park
New York 58, New York

Dear Dr. Rogers:

As I told you on the telephone the other day, I was very pleased to be able to attend your recent seminar and to learn of the very interesting program in taxonomy which you and your colleagues are undertaking.

I find this problem a most fascinating one, and one in which I am personally very much interested, since it has deep roots in pattern recognition.

I was particularly interested to see the approach that you and Dr. Tanimoto have taken.

Since you asked me to jot down any thoughts I might have, I wonder whether you will permit me to suggest a rather wild idea. Would it be possible to develop an automatic taxonomic system which could modify its behavior based on observing how you and your colleagues actually classify. As you know, the perceptron of Rosenblatt is at least a primitive system which does something of this sort.

With very best regards.

Sincerely yours,

Herma Green

P. S. Dr. Alan Konheim of our staff, who has done work on the Perceptron, would be glad to come down at some mutually convenient time to discuss this approach with you if you desire.

AIBS AMERICAN INSTITUTE OF BIOLOGICAL SCIENCES
HOBART 2-6556 2000 P STREET, N.W. WASHINGTON, D. C., 20036

June 26, 1964

Dr. David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx Park
New York 58, New York

Dear Dave:

Although your letter of June 5th caught me in the process of developing a memorandum to you re the meeting, it has taken me a considerably longer period of time to complete it than originally anticipated.

I might add that some of the comments will be personal, others from side conversations with various attendees.

First of all, it was my distinct impression that most everyone agreed with the use of and need for implementation of computers in biological endeavors. As one might expect, there were a variety of opinions expressed as how such usage could be effected.

Some expressed the opinion that current programming was being designed to fit the present capabilities of the instrumentation. No doubt this opinion has validity since such limitations must have been part of your thinking in your initial trials.

Others felt that the programming should be designed first without and strictures imposed by current computer limitations and then have an instrument designed or altered to perform for your needs.

There were many expressions as to how the information should be programmed to give the desired results. I am not a mathematician as such but I had the feeling that not everyone was in agreement with the methods employed initially to provide the data being retrieved. Perhaps such arguments are valid but I will have to leave this area to the slide rule boys.

I heard several comments to the effect that perhaps not enough of the other similar type projects had been investigated. I cannot comment on this but pass it on for what it is worth.

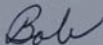
June 26, 1964

There were expressions that the programming was too sophisticated and all that need be done would be to feed the computer enough information with which to make comparisons with unknown entities and/or to provide for rapid retrieval of information on a more general level. These expressions are subject to detailed thought and examination since the program of computer use varies according to investigator, institution and need.

In summary, it seemed to me that (1) computer programming should be utilized in biological investigations, (2) past and on going programs should be investigated to gain experience and prevent duplication of effort, (3) investigations and/or usage of computer capabilities should be programmed step-wise beginning with simple tasks and gradually expanding to more sophisticated ones, (4) several types of programming should be attempted and (5) perhaps cooperative programming endeavors should be examined with other research centers.

I hope this rambling of mine will be of some help to you and your co-workers.

Sincerely yours,

A handwritten signature in cursive script, appearing to read "Bob", is written above the typed name.

Robert S. Leisner, Head
Special Projects



DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

PUBLIC HEALTH SERVICE

BETHESDA, MD. 20014

NATIONAL INSTITUTES OF HEALTH
Area Code 301 Tdl: 656-4000

June 25, 1964

Re: NIGMS-RGB
GM 11208

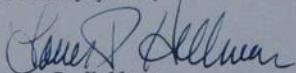
Dr. David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx Park
New York 58, New York

Dear Dr. Rogers:

In your letter of June 5, 1964 you asked for my impressions of the seminar that was held at the Botanical Gardens on June 2nd. Let me say at the outset that I think the program was well planned and that the presentations were both interesting and informative, and indicated to me the high quality of work that is being done at the Botanical Gardens and the great potential for applying the methods of numerical taxonomy. Although we did not do justice to many of the questions that you had planned to have discussed at the seminar, I think that our informal discussions at luncheon and dinner while visiting Dr. Steere and Dr. Fleming brought out the various ways in which your program can be augmented. As I stated then, some of the life sciences areas in which numerical taxonomy can be useful are genetics, pharmacognosy, pattern recognition (bioengineering), psychology, ecology, and medical diagnosis. However, as I pointed out at that time, these problems are most efficiently attacked on a multidisciplinary basis. Thus, I still feel, although many of those at the conference were inclined to disagree with me, that it is necessary to have those in other disciplines interested, and knowledgeable to some extent, in the methods of numerical taxonomy. Thus, a pharmacologist should have some knowledge of what goes on in numerical taxonomy and by the same token, someone on your staff should be knowledgeable in the area of pharmacology. In this way, both investigators can work together and make effective contributions. Along these lines, I would suggest that you encourage students from surrounding universities to come and study your work at the Botanical Gardens and to assist you in your research. Since Dr. Steere is on the faculty at Columbia University, this plan might start effectively with students in the graduate school at Columbia University. As I informed you, NIH would be willing to consider support of collaborative endeavors including staff and the required hardware if it can be fully justified. However, at the present stage of the development of your research program, it does not seem likely that you could justify the purchase or rental of a computer.

I hope that these comments will be useful to you. If they raise additional questions, please do not hesitate to call on me. Again, may I state that I enjoyed attending the conference and the gracious hospitality shown by Dr. Steere and Dr. Fleming. My best regards to them.

Sincerely yours,

A handwritten signature in dark ink, appearing to read "Louis P. Hellman". The signature is fluid and cursive, with the first name "Louis" and last name "Hellman" clearly distinguishable.

Louis P. Hellman
Acting Head, Biophysical Sciences Section
Research Grants Branch
National Institute of General Medical
Sciences



DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
WASHINGTON, D. C. 20360

IN REPLY REFER TO

CNR:AA6:SRG:pc
26 June 1964

Dr. David J. Rogers
The New York Botanical Garden
Bronx Park
New York 58, New York

Dear Dave:

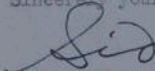
I am in receipt of your two letters dated 5 June and 16 June 1964. Before I begin let me take this opportunity to express my appreciation for being permitted to attend your very interesting seminar and also to apologize for my delay in sending you this note. As you probably guessed, I have been away on a series of field trips which has resulted in the usual delay in acknowledgment of my correspondence.

In response to your request of 5 June 1964, I am presenting the following thoughts for what they are worth. However, I do wish to caution you that they are not intended as suggestions but rather as questions for consideration. Firstly, I came away from the seminar with the feeling that many of the participants in that meeting, especially the mathematicians, were of the opinion that it was essential that you obtain the full time service of a competent mathematician. Was that your feeling also? Certainly I am greatly impressed with Dr. Tanimoto and I think that if you could have him associated on a full time basis with your project your mathematician problem would be solved. However, since this is an unlikely event you may wish to give serious consideration to the mathematician's feelings on the subject. Secondly, I raised the question at the end of the seminar as to the desirability of establishing an advisory group to your project that might serve both as a scientific and technological "sounding board" as well as a missionary group by presenting your views to the scientific public--especially taxonomists. Have you given any further thought to this question? Finally Dave I raised a rather basic question which I hope will not place you under any misapprehension regarding my confidence in your project. The question is simply this; are you as convinced today as you have been in the past that the algae represents the most appropriate group to serve as the "guinea pig" for your model? I raised this question only because of a number of comments which I picked up during the course of the seminar which collectively suggested that the algae were a highly complex and somewhat controversial group from the taxonomists point of view.

Perhaps more to your point of the question is the lack of a full time algal taxonomist for your project. Of course, I realize that you have been seeking support for the hiring of subject person and have not received any great encouragement in this regard. By the same token I would very much like to encourage other groups to join in the financial sponsorship of this very worthy scientific activity. The questions that I have put to you are ones that are likely to be encountered in seeking support from other groups.

In closing let me express once again my appreciation to you and Bill Steere for the initiative which you fellows have shown in convening a seminar without outside support. I wish you the fullest success.

Sincerely yours,



S. R. CALIER
Head, Biology Branch

Copy to:
ONR/New York

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
WASHINGTON D.C. 20333



SRI

22 June 1964

David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx Park
New York 58, New York

Dear Dave,

I feel even more strongly now that there is some connection between information retrieval and taxonomy, even though I am not at all sure what the document equivalent of a "taxon" is. I feel also that if the ratio between the funds spent on information retrieval and taxonomy had been reversed that we might have even more to learn from your field.

I can make no suggestions however as to how to break through your present trap--if you had a computer, you could undoubtedly find many uses for it and if you could justify it, you might conceivably be able to get a computer. I agree completely with Minna Rees that the Rogers-Tanimoto approach is a good, but not necessarily unique nor even optimal approach. It does work, and as long as you are working with minimal resources, perhaps that is all one could ask for.

I was interested, intellectually, in the problems of automating an herbarium. What you now have is a good, perhaps one of the best, hierarchical classification system. What you need, in addition, is a coordinate classification system, cutting across genera--what plants concentrate germanium in their tissues, for instance. It would be hard to justify economically, but it would be fun to try.

This might also be an approach to solving one of your other problems. I would not be surprised to find that the classical botanists at the Gardens regard quantitative taxonomy and computers as an interesting novelty, having nothing to do with their real business. This is

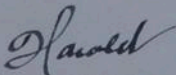
David J. Rogers

22 June 1964

probably not so, but how do you educate them? And, if you don't, how are you going to get a wide base of support for the need for a computer?

Forgive these rambling thoughts, I enjoyed the meeting, more than any I have attended in the past few years.

Sincerely,

A handwritten signature in cursive script, appearing to read "Harold".

HAROLD WOOSTER
Director
Information Sciences Directorate

The Rockefeller Foundation

111 WEST 50th STREET, NEW YORK 20

AGRICULTURAL SCIENCES

CABLE: ROCKFOUND, NEW YORK
TELEPHONE: COLUMBUS 9-8100

June 17, 1964

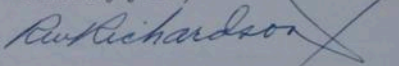
Dear Dr. Rogers:

Thank you for your kind letter of June 5, and also for the opportunity to participate in the recent Seminar at the Botanical Garden. You certainly brought together a capable group of mathematicians. I am sure the dialogue was as helpful to them as it was to some of us in the plant breeding field.

I don't believe I have any further thoughts which will be of particular value to you and colleagues, except to re-emphasize my own feeling that statistical approaches to quantitative inheritance could, I believe, be more fully utilized in solving some of your taxonomic problems than they have been to date. I might add that I was somewhat surprised that the mathematicians present did not spend a significant amount of time discussing this possible avenue of approach. Perhaps their more recently developed formulae are more effective, although I did not gain this impression from the discussion.

By all means this type of open discussion between the mathematicians and the biologists is stimulating and profitable and should be repeated with some frequency. I hope that the Botanical Garden may in the future present this type of Seminar again.

Sincerely yours,



R. W. Richardson, Jr.
Assistant Director

Dr. David J. Rogers
Curator of Quantitative Taxonomy
The New York Botanical Garden
Bronx Park, New York 58,
New York

RWR:lnr

June 18, 1964

Dr. David D. Keck
The National Science Foundation
Washington, D.C. 20550

Dear Dave:

Enclosed are the transcript of your discussion at our seminar and my attempt to summarize your statements. Would you mind examining the two, and deciding whether my reading of your remarks are satisfactory to be included in the final report, or whether you want to reword your own discussion. My only intent in summarizing your remarks is to save your time. If I have done it an injustice, please so indicate.

Your comments in your letter of the 16th were received this morning. I think that your comments indicate very well the problems we have in communicating between taxonomists and others. You indicated disappointment that no more discussion was made about Adansonian methods of using non-weighted information versus the usual taxonomic method of weighting. Tanimoto's discussion, couched in mathematical terminology, did indeed talk about the possibilities, methods, and value of weighting. Taffee and I both feel that we must weight our characteristics if they are to have meaning, and that although Sokal claims not to use weights (Adansonian), he unconsciously does. Henry and I indicated in the rough draft of the paper we sent you before the seminar that, in any system, by any taxonomist, the process begins without weighting, but as the classification develops, certain data are found to be more significant than other pieces of information, and we select (weight) the information useful for the purpose of classifying. But our question is "How do you determine what and how to weight?" If we can answer this question, and I think we can, we will put the process on a much more scientific basis than it presently is.

We hope that we can continue to have discussions similar to the one we held here. There are a number of places that have developed interests in mathematics and biology (with emphasis on the taxonomic aspects), and these will generate much more interest in this vital area. You may be familiar with the conference on microbial classification to be held in Canada this August. It is financed by NSF, and Henry and I have been invited to participate. While the title says microbial classification, the discussion will center around classifications in general and the use of mathematical models of taxonomy.

Thanks for your contributions--we hope to have a full report on the seminar soon.

Sincerely,

David J. Rogers

NATIONAL SCIENCE FOUNDATION

WASHINGTON, D.C. 20550

June 16, 1964

Dr. David J. Rogers
New York Botanical Garden
Bronx Park
New York 58, New York

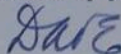
Dear Dave:

I much enjoyed my visit to the Botanical Garden to participate in your seminar. Thank you for your letter of June 5 about it. It was an interesting experience to be in a group of mathematicians and biologists who were trying to learn something from each other. I was hoping something more would come out at the meeting in regard to the soundness of using Adansonian numbers in the classification of higher organisms. This has not been properly established to my satisfaction, although I judge Michener's bees may have been a rather good example. I have not studied his paper with Sokal in detail. At all events in that example the computer was used independently to work with relationship correlations on organisms that had already been thoroughly worked over by conventional taxonomic means. The computer was given data already determined to be significant by conventional taxonomic means - that is, Michener's taxonomic characters were used in the programming, were they not? My feeling is that computer science is perfectly adequate to establish correlations or cross-correlations in any combination desired, but the finding of characters to program still rests with the scientist and not all scientists see the same things. In other words, I think we will be obliged to stay with the conventional methods at the time being in all but the most exceptional cases.

Your pioneering work is interesting, and I see you are going to add much precision to it as you go along. I shall certainly follow your activities with interest.

Thanks again for your fine hospitality the other day.

Sincerely yours,



David D. Keck
Deputy Division Director
Biological and Medical Sciences

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
WASHINGTON D.C. 20333



9 June 1964

Dr. David J. Rogers
New York Botanical Gardens
Bronx Park
New York 58, New York

Dear Dr. Rogers

I felt that the seminar discussions were very informative.

It appears to me (I am no taxonomist) that the fundamental question is really what is desired in a taxonomic system. By way of illustration, all present in the room were *Homo sapiens*. However, if other than the usual taxonomic criteria were used, say finger prints, no two of us were alike. Hence there are always more or less easily identifiable criteria for breaking into subspecies. There is no logical limit that I can see in the "splitters".

I believe this is a philosophical question which will not yield to computers or for that matter, to human thought. However, arbitrary criteria can be agreed upon and imposed.

I still hope that you will consider the use of discriminant analysis. In some fields of anthropology it is proving very useful. It can still require access to a computer in some of the more complex questions.

Sincerely

Robert V. Brown

ROBERT V. BROWN, Chief
Biological Sciences Division
Directorate of Life Sciences

THE FORD FOUNDATION
477 MADISON AVENUE
NEW YORK, NEW YORK 10022

PROGRAM IN SCIENCE
AND ENGINEERING

June 16, 1964

David J. Rogers
Curator of Quantitative Taxonomy
New York Botanical Garden
Bronx Park
New York 58, N.Y.

Dear Dr. Rogers:

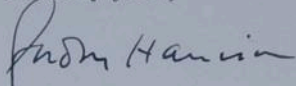
I was very happy to attend your recent seminar and feel that I have got a lot out of it. The idea of trying to develop mathematical language to express and deal with various biological problems is one that will continue to interest us and one that I hope to explore further in the months ahead.

At this moment, however, it is impossible to say what projects, if any, in this general field may seem sufficiently germane to my conservation program objectives to warrant a recommendation for support. As I told you, I do not think it likely that we would want to move in this direction for some time to come and not at least until we have a core program well launched and can define special needs much more clearly than is possible now.

Sorry to sound so negative, but I hope you will understand the problem.

Best wishes,

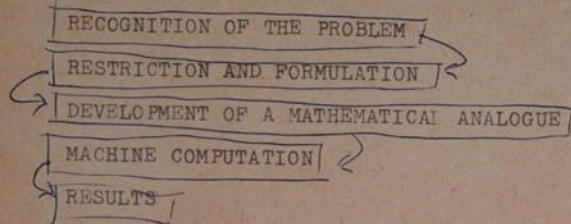
Sincerely yours,



Gordon Harrison
Associate Director

I AM GOING TO DISCUSS THE IDEA OF FORMULATING A BIOLOGICAL PROBLEM
WITH ENOUGH OBJECTIVE PRECISION THAT A MATHEMATICAL ANALOGUE
TO THE PROBLEM CAN BE ESTABLISHED| BY MEANS OF THIS ANALOGUE
THE SPEED→ EFFICIENCY AND HIGH DATA CAPACITY OF TODAY'S ELECTRONIC
COMPUTING MACHINES CAN BE USED TO HELP SOLVE THE PROBLEM| IN RECENT
YEARS SUCH MATHEMATICAL MODELS HAVE BEEN USED [↓] EVEN TO ADDRESS
PROBLEMS OF APPARENT NONMATHEMATICAL NATURE WITH ENCOURAGING ~~SV~~
SUCCESS| I WILL DISCUSS THE METHOD OF MATHEMATICAL MODELING
IN BIOLOGY ILLUSTRATING THE DISCUSSION WITH A PROBLEM ON
WHICH I HAVE BEEN WORKING RECENTLY| THIS IS THE PROBLEM OF
INDUCING A LIKELY EVOLUTIONARY ^{history} ~~FOR~~ SOME GROUP OF RELATED ORGAN-
ISMS| THIS PROBLEM HAS INTERESTED STUDENTS OF PHYLOGENY FOR
MANY YEARS| ^{AND it also makes} ~~THIS PROBLEM MAKES~~ A GOOD EXAMPLE AS IT IS ~~ONE~~ OF
APPARENT NONMATHEMATICAL NATURE|
THE BASIC PROCESS OF MATHEMATICAL MODELING CAN BE THOUGHT OF

WITH THE HELP OF THIS FLO-DIAGRAMME:



4 min

AS WE PRESENTLY RECOGNISE THE PROBLEM TO BE ONE OF INDUCING
 EVOLUTIONARY HISTORY THE NEXT STEP IN BUILDING A MATHEMATICAL
 MODEL IS TO RESTRICT THE SCOPE OF THE PROBLEM TO A WELL DEFINED
 SET OF ASSUMPTIONS ABOUT IT IN ORDER THAT A PRODUCTIVE MATHEMATICAL
 ANALOGUE MIGHT BE BUILT| THEN WITHIN THE LIMITS OF THESE

RESTRICTIONS THE PROBLEM MUST BE FORMULATED PRECISELY IN TO
 COMPLETELY ELIMINATE AMBIGUITY| ONE PRECISE FORMULATION OF
 THE EVOLUTION PROBLEM IS ACCREDITABLE TO JOSEPH CAMIN AND ROBERT
 SOKAL BOTH OF THE DEPARTMENT OF ENTOMOLOGY UNIV OF KAN| THEY
 PRESENTED THEIR FORMULATION IN EVOLUTION VOL LIX 1965.

WHEN I WORKED ON THIS PROBLEM I FORMULATED IT IN SOMEWHAT MORE
 GENERAL TERMS AS FOLLOWS:

AN EVOLVING ORGANISM CAN BE THOUGHT OF AS A NUMBER OF DISTINCT
 FEATURES EACH OF WHICH EVOLVES INDEPENDENTLY OF THE OTHERS|
 EACH OF THESE FEATURES HAS DIFFERENT OBSERVABLE FORMS IT CAN

TAKE STARTING WITH SOME MOST PRIMITIVE FORM AND CONTINUING
 TO MORE ADVANCED FORMS ACCORDING TO SOME SCHEME| THESE FEATURES
 I WILL CALL CHARACTERS AND THE SEVERAL FORMS THAT ANY GIVEN FEATURE
 CAN ASSUME I WILL CALL THE STATES OF THAT CHARACTER| THE SCHEME

ACCORDING TO WHICH THE CHARACTER STATES EVOLVED FROM ONE ANOTHER|

6 MIN

IS KNOWN AS A CHARACTER STATE TREE AND CAN BE REPRESENTED

AS FOLLOWS|

0 1 2 3 4 5

0 0 0 2 2 1

in this way an organism under consideration CAN BE REPRESENTED

BY A SEQUENCE OF CHARACTER STATES ONE FROM EACH CHARACTER USED

TO DESCRIBE THE EVOLUTIONARY DEVELOPMENT OF THE ORGANISMS IN

QUESTION| IT IS POSSIBLE FOR ONE ORGANISM TO HAVE EVOLVED

INTO

~~FROM~~ ANOTHER ONLY WHEN EACH OF THE STATES EXHIBITED BY THE

INTO

FIRST IS THE SAME AS OR COULD HAVE EVOLVED ~~FROM~~ THE CORRESPONDING

STATE OF THE SECOND| A SINGLE EVOLUTIONARY EVENT IS CONSIDERED

TO BE THE CHANGE FROM ONE CHARACTER STATE TO THE NEXT MOST

ADVANCED STATE| THE NUMBER OF EVOLUTIONARY EVENTS REQUIRED TO

EVOLVE ONE ORGANISM INTO ANOTHER IS THUS THE NUMBER OF EVOLUTIONARY

EVENTS REQUIRED TO TRANSFORM THE PRIMITIVE ORGANISM INTO THE

MORE ADVANCED|ONE|| FOR EXAMPLE

1211 → 3212

1→2→3

a tree or EVOLUTIONARY HISTORY ~~OR~~ FOR A GROUP OF ORGANISMS

IS A COLLECTION OF ORGANISMS [SOME OF WHICH MAY BE HYPOTHETICAL

ANCESTORS] TOGETHER WITH A RULE WHICH ASSOCIATES WITH EACH

9 Min^{lv}

ORGANISM IN THE TREE SOME OTHER ORGANISM IN THE TREE FROM WHICH
IT EVOLVED| A TREE OF COURSE MUST CONTAIN ALL OF THE ORGANISMS
BEING STUDY^{red}| AND IT MUST HAVE A SINGLE MOST PRIMITIVE MEMBER|

NOW IF WE FURTHER RESTRICT THE PROBLEM WITH THE ASSUMPTION
THAT SUCCESSFUL EVOLUTION IS IN GENERAL AN EFFICIENT PROCESS

A MOST LIKELY EVOLUTIONARY SCHEME FOR A GROUP OF ORGANISMS
WOULD BE THAT TREE WHICH CAN EVOLVE FROM ITS UNIQUE COMMON
ANCESTOR WITH THE FEWEST NUMBER OF EVOLUTIONARY STEPS| AT THIS
POINT WE HAVE THE PROBLEM SUFFICIENTLY RESTRICTED ^{and have} ~~THAT WE COULD~~
PRECISELY DEFINE IT|

NOW WE ARE READY FOR THE THIRD AND FOURTH STEPS| IT IS HERE WHERE
MY APPROACH DIFFERS ^{significantly} ~~RADICALLY~~ FROM THAT EMPLOYED BY CAMIN AND
SOKAL EARLIER MENTIONED| I WILL NOT GO INTO THE MATHEMATICAL
DETAILS OF THE MATHEMATICAL ~~FOR~~ ANALOGUES FOR THIS PROBLEM
AS THEY ARE NOT OF PARTICULAR INTEREST TO US IN THE TIME ALLOTTED|

HOWEVER THEIR ARE TWO CONCEPTS FROM THE FIELD OF COMPUTING WHICH
CAN BE USED TO COMPARE THE APPROACHES| THE COMPUTER PROGRAM
USED BY CAMIN AND SOKAL IS KNOWN AS A HEURISTIC| MINE IS KNOWN
AS AN ALGORITHM| THE DIFFERENCE IS THAT A HEURISTIC AS THE

MANE IMPLIES IS A PROCESS THE RESULTS OF WHICH TEACH YOU SOME-

11 Min

THING ABOUT YOUR PROBLEM AND MAY OR MAY NOT SOLVE IT | AN ALGORITHM

IS A PROCEDURE WHICH LEADS YOU TO THE ANSWER |

THE RESULTS OF THIS PROCESS ARE INDEED A SUGGESTED PHYLOGENY

FOR THE ORGANISMS IN QUESTION | THE TREE OFFERED BY THE MACHINE

WILL BE THE MOST EFFICIENT ONE |

AN EXAMPLE OF ACTUAL ^{BACKIT} ~~COMPUTED~~ FROM THE IBM 704 ^{ON WHICH THE}

^{can be} PROGRAM WILL RUN, TAKEN FROM A HYPOTHETICAL STUDY ~~LOOKS LIKE~~ ^{what I have seen}

6 5 1

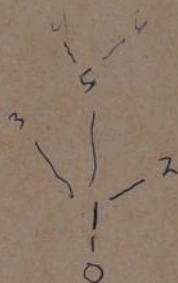
4 5 2

3 1 4

5 1 2

2 1 3

1 0 1



~~Let me quote Camm and Schubert in an analytical~~

15 MIN.

UNIVERSITY OF TORONTO
BOTANY DEPARTMENT SEMINAR

Dr. David J. Rogers
New York Botanical Garden

will speak on

"SOME DYNAMICS OF SYSTEMATIC BOTANY:
THE USE OF COMPUTERS IN TAXONOMY"

Time: 5:00 p.m., Thursday, April 8th

Place: Room 203 (Seminar Room) Botany Building

Tea at 4:30
to meet the speaker

EVERYONE WELCOME

Notes for talk to Biol. group -

1. A short introduction on the nature and purposes of taxonomy - the science, vs. classification, the activity.
2. Taxonomy a way of thought, with its own rules + logic -
3. Classifications are of varying sorts
 1. Traditionally + presently current is intended to make an hierarchical ordering which reflects the evolution.
 2. May ~~be~~ be of special types - genetic, physiologic, ecologic or others - no reasons to confine them.
4. The taxonomic rules - or logic is still not apparent, but some of the procedural steps in classification are becoming defined -
Look at the steps - problem, ~~delimitation~~
→ data gathering → similarities → clustering → formal classification
→ keys, descriptive states.
5. Character selection and the breaking into states - ordering, if possible, no info.
6. Similarity measure - may be a ratio, or a simple matching. Various measures
7. Clustering - graph procedure
illustration - do not be misled by illusion of dimensionality on blackboard.
8. Not an evolutionary model yet.

75d

Assembled one of the best groups of
systematists knowledgeable of computer
methods - including mathematical -
want to

Similarities

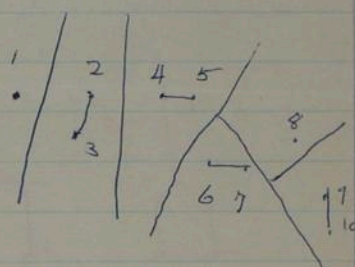
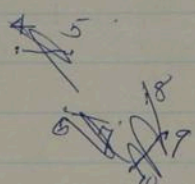
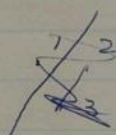
L. 1 - .95

2-3

4-5

6-7

9-10



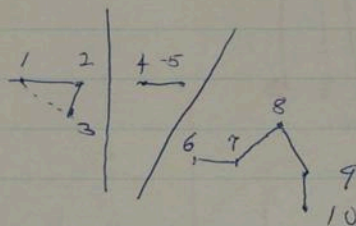
L. 2 - .85

1-2

8-9

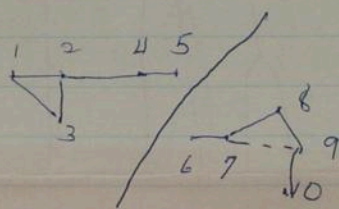
1-3

78

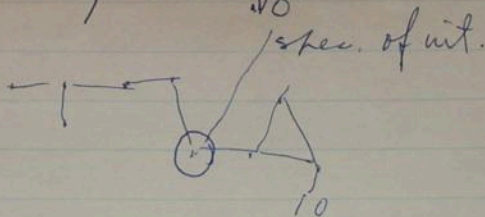


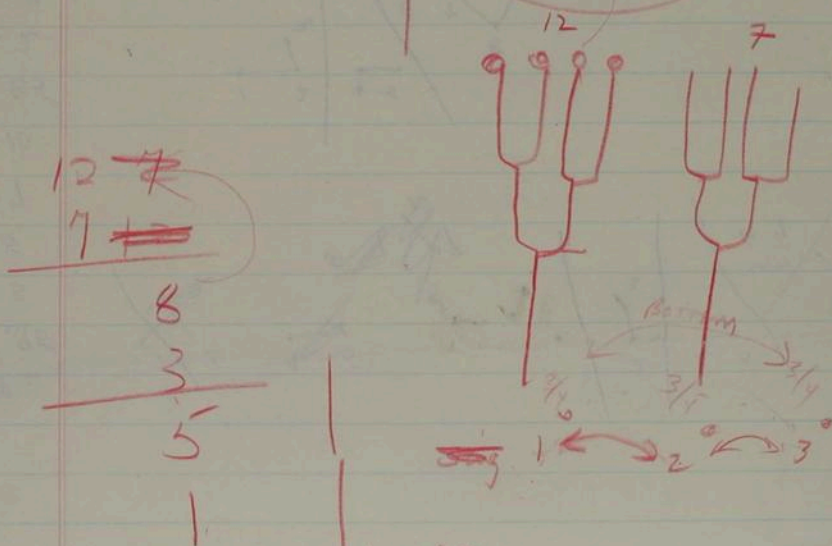
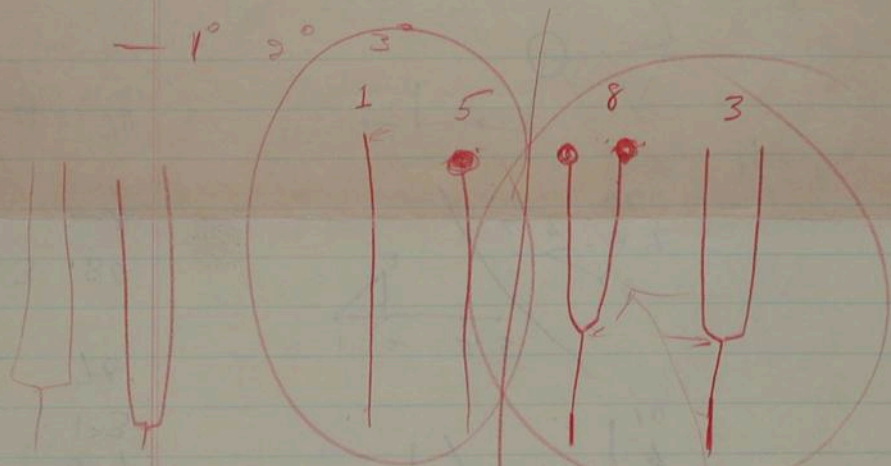
L. 3 .80

2-4

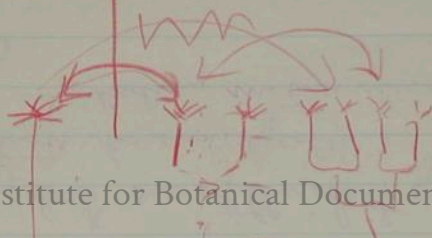


L. 4 .70





None
top
(3/4)
2



THE NEW YORK BOTANICAL GARDEN
BRONX • NEW YORK 10458 LU 4-8500

File Copy

W

July 17, 1964

Dear

Enclosed is a copy of the summary of the June 2 seminar at The New York Botanical Garden on mathematics and biology. We trust that we have a sense of the meeting here and that we have not done an injustice to anyone's presentation. If there are any major errors, would you please let me know?

Once again for the Botanical Garden, thank you for your participation.

Sincerely yours

David J. Rogers

David J. Rogers
Curator of Quantitative Taxonomy

DJR:MDF
Enclosure

SEMINAR PARTICIPANTS
June 2, 1964

Dr. R. V. Brown
Air Force Office of Sci. Res.

Dr. M. R. Carriker
Marine Biol. Lab., Woods Hole

Dr. Miles Conrad
Biological Abstracts

Mr. Edward Fitzgerald
Control Data Corporation

Dr. John E. Flynn
Office of Naval Research

Dr. Sidney Galler
Office of Naval Research

Dr. Herman H. Goldstine
International Business Machines

Dr. Gordon Harrison
The Ford Foundation

Mr. Louis P. Hellman
National Institutes of Health

Dr. Walter Hodge
National Science Foundation

Dr. David Keck
National Science Foundation

Dr. John Laderman
Office of Naval Research

Dr. Robert S. Leisner
American Inst. of Biol. Sci.

Dr. Josiah Macy
Albert Einstein Coll. of Medicine

Mr. Samuel Matsa
International Business Machines

Dr. Ralph Richardson
The Rockefeller Foundation
*111 W. 54th St.
N.Y. 20, N.Y.*

Dr. Mina Rees
City University of New York

Dr. Otto Schmitt
University of Minnesota

Dr. Taffee T. Tanimoto
Honeywell Corporation

Dr. Talbot H. Waterman
Yale University

Dr. Harold Wooster
Air Force Office of Sci. Res.

NEW YORK BOTANICAL GARDEN PARTICIPANTS

Dr. Arthur Cronquist

Mr. T. H. Everett

Mr. Henry Fleming

Dr. Richard Klein

Mr. R. H. Kolkebeck

Dr. Bassett Maguire

Dr. David J. Rogers

Dr. William C. Steere

Summary of Seminar on Mathematics and Biology
at
The New York Botanical Garden
June 2, 1964

Dr. Steere opened the seminar with a discussion of the kinds of activities participated in by The New York Botanical Garden. He traced the development of botanical gardens in general, first as gardens for royalty, thence into herbal gardens for growing drug plants, thence to government-supported institutions for the people. The New York Botanical Garden, under the leadership of N. L. Britton, established in 1899, received its first materials and library from Columbia's Department of Botany.

The Garden means many things to many people: to the city of New York, it is a park, and the city furnishes funds for the park and housekeeping activities. To others the Garden is a set of public displays of growing plants and museum exhibits. To a wide variety of students, organizations and industries, The Garden is a valuable library, one of the outstanding library collections of plant information in the world. To some botanists, The Garden is a large herbarium of over 3,000,000 specimens, and to others, it is a laboratory of basic physiological and biochemical research. The Garden is a contributor to educational activities of the city, state and nation, from the one or two day "how-to" courses, a two-year course in basic botanical subjects for laymen, a children's course, to courses for, and research direction of, graduate students matriculating in Columbia University's Graduate School. The Garden is a publications organization, with five scientific serial journals, in mycology, higher plant taxonomy, economic botany, memoirs of lengthy monographic studies, and studies of the North American flora. The Garden publishes as well a bimonthly journal of popular information for the layman.

The Garden is one of several quasi-public institutions in New York partially supported by endowment funds with a private board for determining how the funds should be spent. Although we have a number of staff members with professorial standing at Columbia, The Garden is not considered an institution of higher learning by the granting agencies because we do not grant degrees, and we are therefore not eligible for institutional grants from the National Science Foundation. However, much of the research done at The Garden is supported by private and government granting agencies.

A part of the research endeavor is the application of new methods to support The Garden's fundamental research. In this area, there are two activities primarily devoted to these goals. One, a punch card system for recording and retrieving the names of plants, is under the direction of a research associate, Mr. Sydney Gould. The other is a more ambitious program to investigate basic models that can be used for correlative endeavors in biology generally and for which we have established a Department of Quantitative Taxonomy.

It was at the suggestion of Sid Galler that we invited you to this conference to talk about the problems that we have encountered in this latter area. There are many difficult questions to be answered, and we need to have your advice on them. Basically, should we be in this business at all, and if so how shall it be financed?

Rogers: Mathematical models for taxonomic work have been attempted for a very long time. Most have been of a statistical nature and have not been successful reflections of taxonomy.

We are interested in the potentialities of useful models and computer programs for taxonomy because the majority of The Garden's research endeavor is centered on taxonomy. If ways may be determined to make practical mathematical and computer models of the work done intuitively by the taxonomist, we will have provided a basic framework for many important studies ranging from genetics to pharmacognosy.

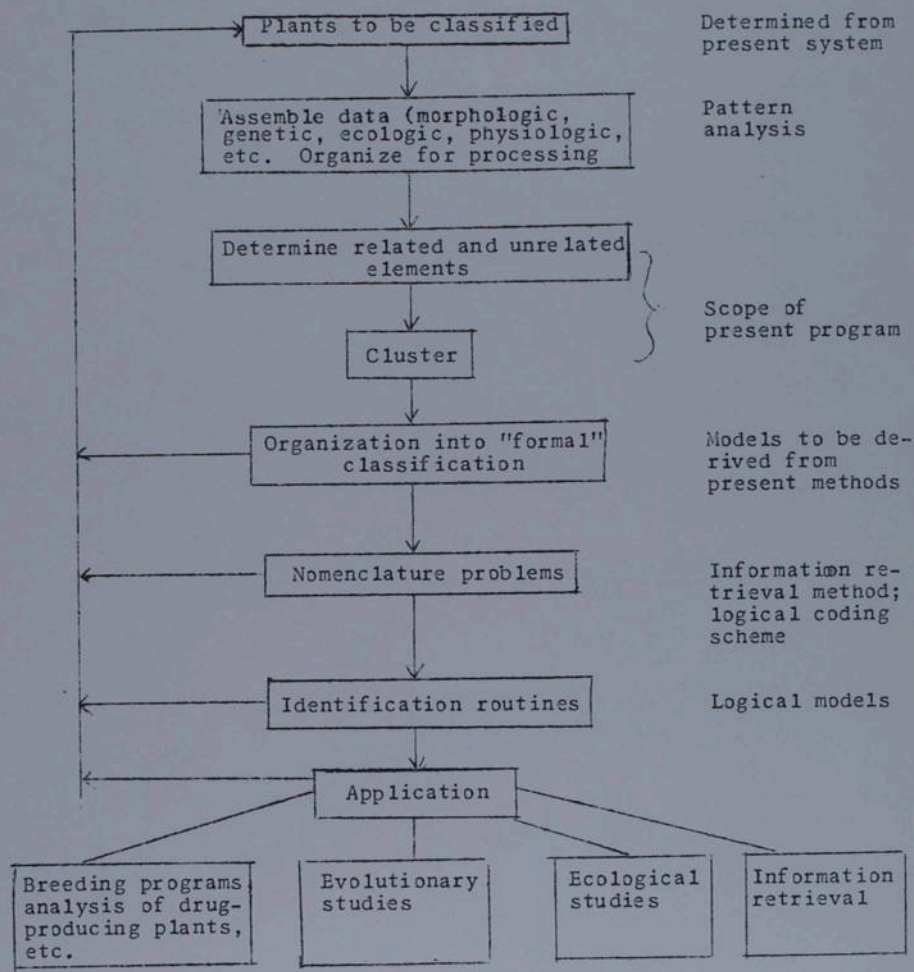
Some of the problems that are amenable to model building and computerization may best be seen if we break the methods employed by taxonomists into their component parts. The flow chart of taxonomy given here is not intended to be complete but rather to give an indication of the regions that we may successfully work on. On the chart I have shown the area that has been studied up to this point--our present model and program cover only one critical aspect of the taxonomist's domain.

The problem, as defined by the taxonomist, may begin when he finds a group of organisms poorly defined and correlated in present classifications. He may discover that an unknown specimen, obviously a member of one group, cannot be adequately placed in the classification. This may lead him to decide to work out a new classification of the group. If this is one entry point to establish the nature of a problem to be solved, we can think of others--a confusion of names for what seems to be one species (a decision reached at the nomenclatural level on the chart). He may use this as the basis for working on a project. The important thing is that, based on present classifications, he may determine the need for further intensive work on one group of organisms. Given the problem, the next step is the selection of material, the decision as to available data concerning the problem group, and the assemblage of needed data not contained in previous classifications of the problem group. Here, the close parallel to the general area of "pattern recognition" may be seen. The taxonomist selects patterns from the specimens representing the members of the problem group that may be compared with other patterns. By making many comparisons of the patterns selected from the specimens, he may be able to provide a satisfactory method of clustering or grouping the specimens into sets, based on the over-all pattern that builds up in his mind. Here, where patterns have been intuitively employed by the taxonomists over a long period of time, we may find a procedure, or model, that reflects the taxonomist's procedures for selecting patterns that adequately represent the real world.

The actual process of the putting the patterns together for purposes of classification of a problem group has been called clustering. Since Fleming's discussion will cover some of the activities involved in our techniques of clustering, I will not take it up here.

Following the clustering procedure, one must make a "formal" presentation of the results. Some ordering of the clusters must be done in order to demonstrate relatedness of the clusters. Following a linear ordering as must be done in publication, one must select, following a

PROCESSING TAXONOMIC PROBLEMS



a strict code of nomenclature, the appropriate Latin names for clusters. The "formal" presentation of a classification incorporates, in a fairly well standardized format, a telegraphic description of each of the clusters (which, when given a botanical name, may be and most frequently is designated as a "taxon"). If one is sufficiently familiar with the taxonomist's method of reporting, he may read in the telegraphic description of a given taxon the defining parameters, the over-all pattern of the taxon, and a number of related pieces of information, both of the taxon, and the habitats where the organism grows. The telegraphic description also permits prediction of additional information. The obvious connection between this type of reporting and model building is the possibility of developing information retrieval methods of heuristic (or) significance greater than we have had to date.

Following the formal classification I have listed, as a separate part of the taxonomic process, "nomenclature problems." These, as indicated above, are difficult to separate from the "formal" classification. However, there are procedures employed in development of correct nomenclature that have little to do with the biological process which went into the clustering techniques. Nomenclatural activities follow a legal procedure, a formal code legislated and adapted by the International Association of Plant Taxonomists. While these rules attempt to give sufficient room for modification of the biological concepts, they also attempt to keep the process of naming organisms within a fairly well defined procedure. Were there no international code, the naming of plants would be chaotic.

As in other areas, laws multiply and become more complex and principles become modified. It is difficult to be certain that in the process of modification, a number of discrepancies occur when one rule is modified, and a second rule is left intact. Frequently there are redundancies in the rules, if not directly opposing statements. Some coding of the rules, with a searching procedure, should allow the presentation of a more logical code for the names of the organisms. This is, however, one of the most complicated types of coding imaginable, and the consideration of varying views of the taxonomists, plus the constant modification of biological concept require much endeavor before the rules can be programmed on the computer.

To many users of taxonomic work, the process of identification of plants or animals is the major part of the taxonomic endeavor. From this chart, one may see that identification procedures are but one part. The identification method employed by taxonomists is familiar to most--keys. The key, a shorthand way of expressing a logical analytical method, is deceptively simple. While it may be employed only as a fast procedure for identification of an unknown, it may more importantly provide the same sort of indication of relationship given by the chemist's periodic table of the elements. The development of a mathematical model of a key, and a program for its use in a computer, should be a potent tool for ready access to information in many areas.

While the processing of a taxonomic problem is essentially complete with the development of a key, we are seldom long satisfied with the work accomplished because modifications and additions to our knowledge of plants continue to accrue from many different areas. For these reasons, the chart indicates a "feed-back" at all levels. If one uses the classification of plants or animals in studies of evolution, one frequently finds information that will affect the classification, and may cause a shift in our concepts of the taxa in the classification. Similarly, ecological studies of the environmental factors affecting the taxa will give additional insights that may or may not change our classification. The constant improvement of classifications is an important work, not liable to reach a static stage. If we are to reflect the most modern thinking about organisms, and the recent investigations of the basic hereditary materials DNA and RNA are providing much insight into the relationships of organisms, we must find new methods to cope with the ever-increasing knowledge. Only by the application of mathematical models, coupled with a large-memory capacity computer, can we hope to successfully fulfill the mission of the taxonomist. It is to these problems that we address ourselves in The New York Botanical Garden.

Dr. Maguire: I have not been attached to this program although it is of interest to me. Dr. Steere has pointed out that the herbarium of The New York Botanical Garden is a large one. It holds about three million specimens. That is a lot of plant material. There are perhaps two other institutions in the U. S. with herbaria to compare with this one in New York. One is in Washington--the National Herbarium--and the other is at Harvard University--the combined Gray Herbarium and Arnold Arboretum Herbarium. Each of these three major institutions has about three million specimens. There are other institutions, many others. Missouri Botanical Garden has about one and a half million, and the Chicago Museum has about two million specimens. Philadelphia has about one million, and many universities, with the exception of Harvard which stands well above all the others, have large and very important herbaria. Altogether then in the major herbaria of the country, there are about 14 million specimens on record that would be of interest to work of this sort. In Europe this process has gone on much longer; there are more herbaria and larger. Perhaps the largest is at Kew with some six million specimens. Paris has five million, Leningrad five million and many have one, two, or three million specimens. The total reservoir runs into the tens of millions of specimens.

The herbarium is a collection of dried, pressed specimens. It is a collection of all kinds of plants. These specimens are arranged in a very efficient system. There are several systems that have developed during the last two or three hundred years which have resulted in the two or three primary systems that apply in the world today. These systems have evolved out of the use, the study and the understanding of relationship of the plants themselves. The system now is so developed that anyone who has a little information or knowledge can put his hand on any specimen that he chooses within just a few minutes.

A specimen is good or bad relative to the care in collecting and also as to the information shown on the label. When anyone goes into the field, he tries to develop the information that was shown on the chart shown by Dr. Rogers: name of plant, the plants that are associated with it, the immediate locality of the plant, the collector's number, and collector's name. It is erroneous to assume that all collected specimens have this information. I wonder how many plants in the herbaria bear the correct name by today's concept--perhaps only 20% of the specimens that we have in here do not bear the correct name. Collections go back further than the beginning of The Garden. Concepts change so what was considered the proper name 50 years ago is not necessarily the proper name now. What will be the accurate name in another 50 years as more and more becomes known about these plants? It means a constant re-working. The older materials that were collected in past generations may have just the one term "Brazil" for the location. Specimens that were collected in the early times in the U. S. would have the single term "Carolina." You have to keep these things constantly in mind. You cannot take off without judgment, information, or data to compensate when information is lacking, merely by a tabulation of these records. The herbarium itself is a tremendously valuable instrument. It is constantly growing in the amount of its usefulness. From it there has been produced a tremendous wealth of books having to do with descriptive characteristics, uses, and the geography of plants. These books are no better or worse than the excellence of the man who has done the work. They are no better than the writer's own interpretation of the material that he has before him. There is an enormous amount of material available. But there still remains some question of interpretation when you use the literature and the herbarium.

Discussion

Wootter questioned the meaning of "type," and whether specimens of each herbarium mentioned were unique.

Replies from Maguire, Cronquist and Rogers indicated that a type is a different concept from the "center of distribution," in that taxonomically, the type is a specimen used to tie the name of a species to a particular item and that the nomenclatural type had no necessary relationship to a "typical" specimen of the species. With relation to the uniqueness (or lack thereof) of specimens from various herbarium collections, it was indicated that parts of the same collection, made by an individual, representing the same plant, or part, could frequently be found in several herbaria simultaneously. That is, quite a bit of duplication occurs, but only when sufficient material may be distributed from one institution to another. Today, collectors will frequently make ten specimens from an individual plant (or of ten individuals in the same locality) and distribute these to as many as ten different herbaria, but earlier collectors may not have had the same desire and collected only one specimen. The present approach is dictated by our shift in emphasis from the individuals to populations.

Rees inquired what happens to cause changes in our concepts of species and what happens in the herbarium when changes are made.

Rogers replied that the investigations made by biologists may indicate that a previous concept of the species was too broad or too narrow. The evidence may come from genetic testing of the species, from physiological analysis of enzyme systems or other functional systems, from more detailed analysis of the fine structures of the plants with such instruments as the electron microscope, etc. More frequently, however, the evidence for a change in concept comes from recent collections of materials not previously studied in the development of the earlier species concepts. The changes made in the herbarium are made by taxonomists who are specialists in a particular group of organisms. The specialists will borrow all the specimens representing his group from this herbarium, and from others, and after his classification has been completed, he annotates the specimens (writes a new label, with his concept of the species), and these new concepts are then followed when the material is reentered in the herbarium.

Schmitt remarked that the fundamental model has been deterministic and inquired whether it should not be more stochastic. (This is an interpretation of the question--question not clearly transcribed at the conference.)

Rogers replied that there are mixtures involved in taxonomic thinking. Mostly we use a deterministic model, but at times, we recognize the stochastic nature, but still work in a deterministic framework. The plants themselves have a fundamental deterministic though flexible mechanism through their genetic content, but in some organisms the expressions possible from the genetic content require a probabilistic approach. Perhaps the species concept is a probability.

Galler offered a suggestion prompted by Dr. Rees' and Dr. Schmitt's questions. He felt that it would be very important to present the concept of taxonomy and systematics, remembering that the group present were not only biologists, in order to get the point of view of the taxonomist across.

Waterman inquired what happens in science. Science is ordered knowledge. He felt that we must recognize that the kinds of endeavor in taxonomy are basic problems to all science.

End of general discussion following
Dr. Maguire's presentation

Fleming: (The first part of Fleming's discussion briefly reviewed the paper sent to all participants.)

The model we have developed is clearly an attempt to give a reflection of the taxonomic process rather than to attempt a new way of thinking about theoretical taxonomy. The most complete test of the model was made with a monograph of a small group of algae prepared in a fashion comparable to the chart given by Rogers. If the model worked, reproducing the same decisions, or nearly the same, as those of the author of the monograph, we would feel successful. In this respect, the model is clearly a success.

The mathematics of the model is, so far as we know, new. That is, the methods employed to reflect the thinking process of the taxonomist have not been applied to taxonomic problems by any other group. What was needed was a method that could be applied to a problem where some sort of pattern existed (or was thought to exist), but because of the numbers of variables, the orderliness was not apparent. Therefore, a statistical model that assumed a random situation was not appropriate.

One of the most interesting aspects of the present model is Tanimoto's adaption of the concept of entropy as a basis for clustering or the determination of sets of individuals. The entropy function is a measure of the homogeneity of a cluster. Formerly, taxonomists have used the designation of "good species" where little variability was found, or "variable species" where many variables were found within the species. The entropy function gives a quantitative measure of these terms.

To date, the method has been employed only for clustering objects. However, one of the major reasons for the development of the method was to test the usefulness of the many pieces of information that define the objects. If we use the program to cluster the attributes of the objects, we will have a very significant contribution to knowledge of the usefulness of the selected information. We will then have more insight as to the values, or weights of characteristics.

We still must make a statistical analysis of the distribution of biological data and the objects to give a better method of determination of the constants employed in the program. At the moment we apply a posteriori methods to decide the cut-off points between the various taxa.

The program is being worked over at the moment to allow the best use of the computer's memory. Several places in the program can be improved and will be in a short time. While this "maximizing" is in process, we will use the program in its present form.

Dr. Tanimoto: An assumption is made that the botanist is able to collect plants that validly typify the real world of plants. We further make the assumption that we can code the properties of these plants by the numerical numbers 1 and 0, in effect, 1 for a property being present and 0 for the absence of the property. The third assumption is that the redundancy or in other words the repetition of properties of these collected objects (plants) will be such that a mathematical weighting of those objects that are similar to one another is such as to lend itself to objective mathematical analysis.

One of the problems which really comes up over and over again is this matter of weighting--i.e., how significant an attribute is, whether you should weight something that is not as important as another attribute, etc. How do you really weight? Do you start from scratch or do you just arbitrarily assign them? You have to start from some point that is very "objective." Basically, all of this is subjective, but one would think that machines are a little more objective. What we do is try to code events or objects by their attributes, and in order to be as impartial as possible, we use 1's and 0's, the binary coding, as the absence and presence of attributes. Whether or not this particular code of 1's and 0's represents the event or object is really questionable. However, there has to be a common language somewhere. The assumption is that the taxonomist goes out in the field and collects. In a sense we base our whole model on his intuitive collecting procedures. Again I don't know how to get around this intuitive aspect, but essentially it is a coding problem. You are representing plants by strings of numbers. For all practical purposes that you have in mind we hope the strings of numbers will be a unique representation. I shall assume that somebody has collected a set of plants and in some sense had ideas of what they were looking for so that they can start to organize them. One thing that one can do is to collect many kinds of plants randomly and then determine the weighting by frequency. This is really operating as if nature were operating. Nature is extremely redundant. The thing one can do is to be extremely redundant and then determine a normalized weighting. Once you have a weighted system and you have a good representation of objects in the universe, then I don't think it is too critical what kind of a taxonomical model you use if it is a reasonable one. What I am trying to state here is that if you have a finite factor representation of any event or object, provided that this is a very efficient representation for your particular problem, then you can apply any reasonable standard statistical model. There are many things you can compute--e.g., simply take components and take the absolute deviations. This gives you a correlation function of the simplest type. What is usually done is a sort of a factor analysis, namely, you look at the co-variance matrix. This is simply the matrix of the usual inner products: the correlation co-efficient. You get a co-variance matrix which is then juggled around and one looks for the principal axes. This is the usual analysis which is associated with a Gaussian type of multivariate analysis. Essentially you are doing a kind of classical analysis that has to do with taking a symmetrically positive definite matrix down to diagonal form. You can do similar things with the absolute deviations. In the case of the ordinary type of analysis you are looking for directions of the principal axis in the pair-wise combination of axes. You can look at triplets and quadruplets of vectors. These go under the head cumulative functions. Now you

can generalize this whole notion. The LaPlace distribution is the simplest case.

$$P(x) = ke^{-\frac{1}{n} |x - \bar{x}|}$$

Goldstine: You are associating with each object a factor with a shortened dimensional space. What are you assigning? Is it the metric to this space. What is the metric?

Tanimoto: The first metric is the absolute difference. The second metric is the quadratic form.

$$xA^{-1}x^t$$

Goldstine: Two metrics in the space?

Tanimoto: As far as taxonomy is concerned, if you want to use these measures as the measure of clustering, what you do is to determine the distributions of this type, put a plane between two clusters and say you find the plane which minimizes the error of classification. If you can generate weights, then you can resort to these usual means.

Goldstine: This is not a mathematical operation but a botanical one. All of the things that you do have to be measured by the degree by which they satisfy the experience of the taxonomist.

Tanimoto: I think that if you are extremely redundant you can pull out certain common notions.

Schmitt: Would you define what you mean by redundancy?

Tanimoto: These are statistical redundancies. The only thing I can say about the model is that had you had these redundancies, then the empirical probabilities of occurrences can be multiplied. You can weight on the basis of this likelihood function. The so-called cluster function is not satisfactorily defined. I don't have much more to say. I would be glad to discuss any aspect of this with anybody.

Goldstine: I don't know how a taxonomist goes about classifying a plant and I would suppose that if one made a mathematical model, he would try to emulate what the expert did. There is something that he does from which we ought to be able to get a great deal of knowledge.

Cronquist: What Dave and his crew are trying to do is to reduce to a mathematical model. He is trying to put this on to formal mathematics.

Goldstine: The proof is whether these techniques produce a classification scheme which a taxonomist would be satisfied with.

Wooster: It is at least debatable as to whether this is retrieval or not. If presumably Tanimoto could come up with a classification scheme which bore no resemblance to what a taxonomist did however done, this might be a better scheme than any taxonomist had come up with.

Goldstine: One has to accept their classifications and see why it works.

Wooster: There is no real grammar for any particular language. Similarly taxonomy is defined as that which the taxonomist can remember.

Macy: One of the unanswered questions is the criteria for knowing when you have as good or better answer.

Cronquist: The present scheme could not be completely at cross purposes with what ought to be the best. Any two things that can cross and produce offspring must be pretty closely related. There is no known example of a hybridization between things which are regularly put into different plant families. As soon as you get to the family level, there are no hybrids. There are some at the genus level. If the scheme were set up wrongly, we would have hybrids between families, between classes, etc.

.....
Rogers: Several of the participants have felt that there is a lack of basic information on the nature of taxonomy. We have therefore drafted Dr. Keck to make a generalized statement about the activities in taxonomy over the years and its present thinking. We appreciate the fact that Dr. Keck is willing to enter the discussion without giving him any advanced warning.

Dr. Keck: Attempts at classification of plants have been made for at least 2000 years, and the first efforts were by the Greeks. The needed classifications were quite simple, and served whatever purposes were required for a number of centuries. More or less intensive investigations of the classifications were begun about 200 years ago, and Linnaeus' works were the most significant attempt to make a more orderly system. Though Linnaeus' taxonomy was, by modern standards, completely artificial, his consistent application of a binomial naming system has proved to be one of the most useful techniques. His work stimulated study by many other biologists, and classification of plants was put onto a reasonable, "natural" basis by the middle of the 19th century, before the publication of Darwin's "Origin of Species." It is interesting to note that, though there was no explanation of why the organisms could be placed together in a natural system, they were more or less organized into essentially an evolutionary system before Darwin's publication. This is not intended to indicate that the system established over a century ago was a perfect reflection of our ideas of evolution of plants today, but only that the evolutionary relationship of many of the major groups of plants were already obvious.

In the analysis of a group of plants for purposes of classification, the taxonomist employs information from the plants in whatever form is useful. In many cases, the establishment of a classification is dependent on information subsidiary to the plants themselves. Thus, geographic distribution and habitat information frequently are fundamental, and the knowledge that one group of organisms may have a world-wide distribution, as is the case of the members of the family Rosaceae, may be critical for understanding the taxonomy of a group.

In taxonomy, workers have attempted to develop a precise set of terms, rather than equations, that give a clear picture of the species, genera and family groups. Efforts have been directed toward the reduction of the number of terms necessary to circumscribe the members of the plant kingdom. In the process, a precise format for the descriptive terminology reduces the amount of space necessary to carry sufficient information to determine the limit of variation allowable in one species, or other taxon.

In addition to problems of determination and description of the taxa, the taxonomist is faced with a heavy load of nomenclatural problems. One of these is important in deciding whether a particular organism has been described previously. Many cases of synonymy are known, where the literature is widely scattered and not well documented. We have many situations where, if our legal procedures were followed strictly, we would upset names that have become well established and known by large numbers of workers, and the practicality of the situation demands that we stabilize our nomenclature.

General Discussion

Steere: Calls for consideration of listed questions

Hellman: What actual collaboration do you have with the universities of New York City?

Rogers: The closest arrangement is with the graduate faculty of the Department of Botany at Columbia University. Dr. Steere is a professor, and there are a number of staff members from the Garden who are adjunct professors who teach and function as directors of graduate student research. Graduate students do their research work at The New York Botanical Garden and their degree work at Columbia. We have had similar arrangements with other universities--Fordham and Rutgers, for instance.

Flynn: I would like to throw in an idea. There was some discussion a few minutes ago as to what the taxonomist really does. I think the answer is that he does a number of things in addition to serving as a curator, housekeeping his collection so that scholars can consult it. He first identifies specimens personally unknown to him. This is information retrieval. One of the functions of computer taxonomy will be to reduce if at all possible the drudgery and the hours and days and weeks that may have to be spent in identification of unknowns. The descriptive literature is immense, and its condition chaotic. You may have to wait five or ten years before you get an answer if you send specimens to be identified to an authority of a special group. Second, the taxonomist not merely identifies specimens which he collects. He takes a group, as Dave Rogers has taken *Manihot*, in which the specimens are utterly protean, and he tries to arrange these into species which will be as natural as possible. Now he goes about this depends on several things--what group it is, for example. In certain orders and families of plants hybridization is going on very rapidly; in other groups, mutations. What you deal with is a collection of discrete things, but if you know enough about them, they will be connected. These two situations exist. Ginkgo was mentioned as representative of

the discrete type of thing where hybridization is not a factor, where intermediates do not present problems. The hawthorn group was mentioned and is a good illustration of the opposite instance. Procedures are of one kind when he has things with few or no intermediates and a different procedure when he is working with things like the hawthorns, so protean that they almost defy description. What he does depends to a degree on his own interests and training. Paterson used a genetic approach; Anderson, with a different background, another. What procedures he follows differs greatly from man to man according to his estimate of the importance of the geographic variability, hybridization, genetic factors and a great many other things that it would be foolish for me to try to describe this afternoon. The use of cluster techniques, of the computer, and of statistical studies and the like can be used to simplify taxonomic procedure.

Over the two centuries that separate us from Linnaeus, it seems to me that there has been a thoroughly consistent but not uniform effort on the part of taxonomists to get away from subjective and intuitive classification and to substitute objective for subjective methods wherever such substitutions seem to be possible. Different scientists working at different times and in different nations have contributed individual insights and sometimes useful insights which lead to improvement. Charles Darwin introduced one of the biggest insights which changed the whole tenor and purpose of taxonomic study. There have been others, a great many others, whose contribution to taxonomy has not been primarily a classification of any particular group of plants or animals beyond elaboration of a new way of looking at things.

Discussion following Dr. Flynn: A number of points were brought up from the floor, relating to various aspects of the problems of computer use, of the difficulty of language communication, of the nature of information retrieval as used by taxonomists, of various mathematical techniques--e.g., discriminate analysis for taxonomy, the types of mathematicians required and the possible methods of collaboration between biologists and mathematicians. Any one of these topics could be the focal point of discussion of a full day's symposium, and they indicate the need for airing of the problems. Dr. Galler suggested that there be a continuing discussion, with an ad hoc group of mathematicians and botanists to serve as a sounding board for the development of appropriate programs.

Dr. Waterman: The mathematical problems discussed at this seminar are but one part of a much larger problem. One of the difficulties is that many investigators using high-powered tools are not well trained as biologists and that many biologists are not at all mathematically oriented. To help to alleviate the difficulties, a course has been organized at Yale for both theoretical biology and mathematics, with numerous contributors from various mathematical disciplines who have applied mathematics to a wide variety of biological problems.

Dr. Schmitt: There is a surprising lack of communication between two similar efforts, one in medicine, one in taxonomy.

Rogers: We feel that our program will have applicability to medical diagnosis. We have a great advantage in testing our procedures in the plant specimens used as final checks of the procedure. A number of attempts have been made to use our model for various medical diagnostic works.

Dr. Macy: What is really needed is to find individuals interested in the problems and willing to work on them.

Dr. Wooster: There are some people, taxonomic types, who think in pictures, not verbally, and there is no room in present colleges except for the verbal type. Where are the taxonomists coming from?

Dr. Steere: In biology itself, the molecular biologists are hardly on speaking terms with the taxonomist.

Hellman: What other areas can we go into--as many areas as you have people that are knowledgeable in that field to work with a mathematician and a taxonomist.

Rogers: Asks for summary of meeting.

Waterman: Most of the meeting^{was} devoted to a presentation of the various aspects of the botanical garden's interest in a particular problem. In the series presented by the staff those of us who were least familiar with the publications and the problem from other sources were checked out on the particular details of what they would like to do here and what they have been trying to do. There were many cross currents which arose in the course of this discussion and these were very largely generated by the wide range of backgrounds of the people here and by our prejudices and by our imaginations and our interests and unfortunately most of us have been seeing the very important program that the Botanical Garden is trying to get under way through our own glasses so that we see it as related to our own program, to our own philosophy as to what we think biology is. My own reaction is that I would feel that the meeting has been a very satisfactory one in that it has communicated to a lot of potentially interested people the directions of the group here at the Garden. We haven't gotten around to how to pay for it. We have seen that there are strong differences of opinion as to whether you can do this and if this is the right way to do it.

Conrad: My acquaintance with taxonomy is limited, but I came to this meeting more prepared perhaps to discuss the last few questions rather than the earlier ones. At Biological Abstracts, we make use of computers as editorial clerks, not with high level mathematics. This has been an economic boon. I too am concerned as to now this kind of endeavor might be supported. I would appreciate hearing some comments on that from someone now.

Schmitt: We ought not to be asking the question of whether bio-mathematical..... but how and when the first stage is to be done. Secondly, I am a little bit disturbed that we haven't raised the question of generating a mathematics that are appropriate and computer machines that are appropriate in the problems that we face.

Rees: The question of whether you should do it and how-- I feel that you need a clear approach to the problem. You and Dr. Tanimoto have been working together. Mathematicians are pretty narrowly oriented. I don't know how to do it. I think you would want to get a different collection of collaborators. It may be that you are on the best possible track, but how are you going to find out. It would seem to me very important that if you want to make progress you get collaborators that are additional to the present ones. You have got to have people on both sides who want to put something in it. Young mathematicians might have more to offer.

Goldstine: Fordham is getting a computer which means that you have some colleagues across the street.

Rogers: All I can say is that I deeply appreciate your taking a little time out to come and let us express our aspirations here. I hope it has been a profitable discussion for you--it has been for us. It is a beginning of a series of conferences; it is not the end. We are committed to continue--now this ends up we cannot foresee. We hope that we will have at least a sense of this meeting to report to you. Thank you for coming. We hope we will see you again.

Adjourned at 4:30 P.M.



BIOLOGICAL ABSTRACTS

3815 WALNUT STREET PHILADELPHIA 4, PA. U.S.A.

Telephone: EVERGREEN 6-0414 Cable: BIOABS Philadelphia TWX: 215-569-8821

Editorial Offices: PHILADELPHIA • MEXICO CITY

BOARD OF TRUSTEES

WILLIAM C. STEERE
PRESIDENT

FLOYD L. DAFT
VICE-PRESIDENT

KARL H. BEYER, JR.
ROBERT S. CORNELL
D. MILES CONRAD

HERMAN J. EVANS
SECRETARY

ALFRED E. EMERSON
JOHN W. FORD, JR.
N. C. JORDEN
THEODORE L. LEWIN

RAYMOND L. ZWERNER
TREASURER

L. E. PORTER
KENNETH B. RAPER
F. W. WENT

D. MILES CONRAD
DIRECTOR

PHYLLIS V. PARKINS
ASSISTANT DIRECTOR
FOR EDITORIAL AFFAIRS

ROBERT A. GULICK
ASSISTANT DIRECTOR FOR
ADMINISTRATIVE & BUSINESS AFFAIRS

BIOLOGICAL ABSTRACTS

B.A.S.I.C.

Biological AIDS

June 5, 1964

Dear Dave:

I promised that I would write down some thoughts regarding the six questions generated for last Tuesday's most interesting seminar, so here goes.

1. The problems of biology that can be more effectively treated through the use of mathematics --not necessarily "mathematical investigation"-- are certainly myriad. Nor does such use confine itself to statistics. Many years ago, for example, one of my interests was the body forms of fishes --not merely the morphology, but the hydrodynamics of such forms as related to the habits of the several major body-form groups. At that time towing tanks were rare and the only way to test and analyze such matters was in wind tunnels. Although I never became much of an aerodynamacist, I do remember struggling with the mathematics that would help make the observations meaningful. Indeed, my early career in comparative anatomy impinged frequently upon the empirical principles of several other fields of engineering and mathematics was frequently required. In short, in addition to the obvious values of statistical methods in any comparative study, I have found that numerous kinds of biological investigation must involve other branches of mathematics or progress cannot be made. I am not sure, however, that this answers your question regarding mathematical "investigation."

June 5, 1964

2. On the basis of the foregoing, my personal opinion is that biological research of almost any nature, wherever conducted, will always sooner or later find that only through the application of mathematical techniques can one bring orderliness and understanding out of seemingly desperate or chaotic data.
3. I am inclined to agree with several of the participants in the seminar that "mathematicians should remain mathematicians" and that "biologists should remain biologists." They should work together when profitable but should ordinarily stick to their respective lasts. Your association with Tanimoto seems to be the ideal one. There have been some first rate biologists with a happy and erudite faculty for mathematics, but the only one I was ever privileged to know was G. G. Simpson. It's probably too much to hope that such folk can ever be developed in adequate numbers.
- 4 & 5. There is no question in my mind that a computer facility can be justified in a botanical garden such as yours with an active research program and museum responsibilities. The nature and capacity of the computer, however, should naturally be consistent with the tasks it will perform and the volume of such tasks. If a good share of the basic costs could be underwritten by the automation of relatively simple, regularly recurring tasks, there would still be much time available for biological research use of the facility. Among the routine uses might be the obvious ones of bookkeeping, accounting, membership records, billings, analysis of monthly garden attendance; scheduling of garden maintenance, etc. These, when added together, would go a long way to supporting the data processing staff and their basic equipment (savings over older ways being applied to this work through normal staff attrition of the clerical and administrative force). Research involving these facilities and the computer should only be undertaken when funds for "rental" of the facilities are budgeted.

might even be
"desperate"
sometimes!

Further, in view of the state of computer developments, I would recommend that you only consider its rental; outright purchase is fantastically high, as you know, and changes come fast these days.

June 5, 1964

As some sort of experience data for your consideration, BA has been spending \$50,000 to \$60,000 a year on its Data Processing Service. Of this amount, roughly one-half is devoted to personnel salaries and the other half to rental of punch card equipment, sorters, etc. and rental of computer time. For a detailed breakdown of work done ask Bill Steere for our Annual Reports for 1962 and 1963. Even this investment, however, is not sufficient to justify the full-time rental of a small computer at present prices. It is possible that initially your facility could be confined to the in-put staff and equipment with somebody else's computer being rented only for those relatively few hours needed.

6. Here, naturally, I gave some thought and discussion to our mutual sharing of the computer. However, our deadline requirements are so frequent and critical we really do need to be located very close to our computer. Further, because we offset from the print-out directly, we spend much time in "quality control" of the print-out.

There is no doubt that taxonomic lists and the association of generic names with their appropriate higher taxons is a matter of daily concern and interest to us. If your tape records contained this information, we might often make referral to it. Indeed, there must be a number of ways in which an information network among several biological institutions could be linked. For example, BA might tie in with MBL at Woods Hole for animal systematics and the Garden for plants. In turn, your research staff could query us for unit abstract service (this is something we're thinking about now.)

Well, if I don't terminate this soon it won't get in today's mail. However, let's keep thinking and talking, especially on question 6. I would like to hear about your ideas on this one.

With best wishes and many thanks for a most interesting seminar. Hope to see you soon again to bat this around in greater detail.

Sincerely,



G. Miles Conrad
Director

GMC:nmp

Dr. David J. Rogers
The New York Botanical Garden
Bronx Park
New York 58, N. Y.

cc: Dr. William C. Steere

Questions on taximetrics

P. Dansereau

1. Is taxonomy a method of ordination restricted to groups of plants and animals?
2. Can taxonomic methods be validly applied without reference to natural order?
3. What is the basis for selecting criteria:
 - a) if only usefulness is involved?
 - b) if objective relationships are implied?
4. Can computers lead us to the discovery of a hierarchy in characters?

MEMORANDUM 12/19/62

Henry Keep this
somewhere
in a file marked
"Seminars"

To: Dr. Dansereau

From: Dr. Rogers

In accordance with our conversation yesterday afternoon concerning seminars on the computer, I have the following comments which you may wish to circulate. I would like to examine rather informally in about four one-hour periods not only a description of computer hardware (what they are, the kinds of computers, their capacity, speeds, input and output information). I would also like to examine some of the basic ideas of taxonomy which I have learned a little better as a result of trying to use the computers. I would like to speak, for instance, about the potentials inherent in the taxonomic method for a larger range of biological problems than purely taxonomic and to say something of the principles of taxonomy and how hierarchical organization and relationships are valid methods for higher meanings. I would like then to go through the program that we have written for the computer starting in with a description of our program and some of the mathematical ideas involved. I would like to show how the machine operates from the moment of data gathering until the final output. It will be necessary to have a demonstration of the actual computer operation to wind this up, and I hope that we will be able to do this at the New York University Computer Laboratory just off Fordham Road. I will have to arrange this last one with the director of that laboratory who may be disposed to let us come in to the computer rooms for such a demonstration. Lastly, I would like to say how the whole procedure might be and has been used in other than purely taxonomic problems.

I said above that I will try to do this in four times, and I think that I would like to keep it down to that, at least for this first experience with it. Obviously, I do not know how far to go and how much time it will take to make a meaningful presentation. If possible, I would like to keep it flexible so that if we see that we are going ahead faster than I now expect to, then we can cut the presentation time down, or vice versa. If anyone gets excited about the program, then perhaps the whole thing could be extended.

I will ask Henry Fleming to present some of the aspects of this work. I don't think we will be ready to go on this before February. The timing on it is not important; it could start in March if this is more appropriate. I hope that we can find a time that is suitable to all for this activity. You may wish to check with Pat Pirone to see whether or not he would consider this in his bailiwick of seminars or whether we should keep them as a separate organization.

January 3, 1963

MEMO: from P. Dansereau

You will find herewith copy of a proposal by Dr. David Rogers which I hope will interest you. We at the Botanical Garden are in a singularly good position to engage in inter-disciplinary undertakings. There are, in fact, quite a few research projects at this time that involve members of our staff who might seem to occupy opposite ends of the spectrum of biological science.

The development of common (and hopefully unifying) concepts and the utilization of a common methodology can do a great deal to reconcile what seem to be widely divergent aims.

Will you kindly let me know if you are willing to give a few hours to Dr. Rogers' experiment?

January 3, 1963

MEMORANDUM from David J. Rogers

SUBJECT: Seminars on "taximetrics"

I would like to examine rather informally, in about four one-hour periods, some of the basic ideas of taxonomy which I have grasped a little better as a result of trying to use computers. This would include some of the principles of taxonomy, and how hierarchical organization and relationships are valid methods for higher meanings; a description of our computer program and perhaps a demonstration of the computer working on our program. This last item would have to be arranged at the convenience of the Computation Laboratory at NYU where we work.

If it seems fitting, we might add a period in which we examine some of the various aspects of computer hardware -- the various components of machinery, how they work, and their capacity for various operations.

I said above that I will try to cover the subjects in four periods of one hour (including discussion). If we see that we are going ahead faster than I now expect, then we can cut the presentation time down, or vice versa. I will ask Henry Fleming to present some of the aspects of this work.

We don't expect to be ready to go with these seminars before February. The timing is not too important; it could start in March if more appropriate. I hope we can find a suitable time for all interested in this work.

Taximetrics

Seminar, NYBG, April 8, 1963

1. Purpose. An exposition of our experiments in developing computer simulation of taxonomic processes.

a) to help us formulate our own ideas.

b) to stimulate thought along the lines of what the taxonomic method is, and to inquire whether our present mathematical model of the T.M. has any validity.

2. Procedure. Today's discussion lays the ground-work, in part, giving some of the theoretical aspects of the taxonomy, and a start towards the description of our computer program. Next week, Mr. Fleming will give an exposition of the program itself, and the week following will be a visit to the computation lab. of NYU, to see the computer in operation, hopefully with our program.

3. What is the taxonomic method? Hard to define, as all definitions are operational models of what one does taxonomically.

The T.M. is basically similar to other correlative endeavors, such as those in ecology, anthropology, sociology, but the biological T.M. is quite distinct in operation, and in other respects. T.M. encompasses all that is called "classical", "biosystematics" and "cytotaxonomic" methods.

We are not considering an allied, though important, adjunct to the T.M., the nomenclatorial aspects.

The output of the T.M. takes various forms, many incomprehensible to all but the initiated specialist: monographs of varying sizes of taxa; floristic works of differing types (natural geographic areas, or political units), or short notes.

The T.M. includes some interesting procedures: keys, the species and higher category descriptions in telegraphic style. Condensation of information in the descriptions is packed with information, but requires some skill in interpretation. Keys have been put into a number of retrieval formats--the most common being the edge-punched Holorith card. A recent attempt (by Osborne) has been made to use numerical notations for keys. In the computer we actually handle key items, but in a much more rapid means than that by cards, making

The T.M. deals with the classification aspects of taxonomy. Identification of organisms is one of the major uses of the classification.

Definition. Essentially, the T.M. is the process of organizing, ordering and/or arrangement of information about naturally occurring (or cultivated) plants and animals, utilizing largely external traits of the plants or animals.

The purposes of T.M. are various, but in general, attempt to place like plants together, in hierarchies of ascending order, each hierarchy more inclusive than the next lower.

The value of T.M. is that it allows us to think about the individual or aggregations of individuals (populations), and to provide a framework for various types of endeavors--evolutionary, physiological and ecological, genetical and plant breeding, and many others.

We will not argue here about the naturalness of, nor the biological validity of, classification. We assume that our purposes are to give us some logical procedure to grasp the ideas of the universe, and for our own special purposes. The arguments concerning the reflection of nature by a classification are so involved in logic, semantics and philosophy that there is little hope of making definite statements here. Such questions as "is a species a naturally-occurring object?", "do genera and families 'exist' in nature?" may eventually be answered, but the way I see it, until much more data are correlated we cannot even state the questions correctly. The reason² I shy away from this type of question are as follows:

1. The objects that are the subjects of our classification are not studied in toto. Only certain aspects, called characteristics, which serve our taxonomic purposes are selected from the objects for correlation, in isolation from the object.
2. The method of selection of characteristics is largely intuitive. Many times we do not select the characteristics most important for classification-- we do not "see" the characters, or for some unstated reason, we reject them as unimportant. When a character is selected, we may then divide it into different states or expressions. There is most frequently no way to evaluate the genetic importance of a selected character. Questions concerning the number of genes involved in a particular character are not often known. Therefore the biological importance is not assessed.
3. Characteristics of taxonomic ~~infer~~ importance vary from group to group, and we cannot make a one-for-one comparison between the groups. Thus, a species in one genus of one family is defined by morphological characteristics of the flower, in another by the ecological differentiation, and in a third by its incompatibility genetically.
4. Only the broadest type of guide-lines are provided or specified for all organisms, and justifiably so. We may say, after extensive empirical testing, that reproductive structures of plants will provide more reliable characters than will vegetative ones. This does not mean that each classification must consider them in this order.

Our work has been in that section of the taxonomic method beyond the problems of character selection. We assume the validity of the characters, and work on their interrelationship and their value in forming clusters, or groups. The problems of character-selection are in need of much work, however, and are being studied by a large number of people. The behavioral psychologist is perhaps at the fore-front of this type of investigation, but many others are looking at pattern analysis. Even the physicists are beginning to have classification problems with their increasing numbers of particles.

We are in virgin territory as far as mathematical expressions of the taxonomic method is concerned. The scientific models which have been established have been for the analytical sciences, and little has been done for the comparative, synthetical sciences. Most of the endeavors and models that are available are in the realm of statistics. (see Sokal, Sokal and Sneath, Rolfe and Sokal). Some parts of the taxonomic method are indeed statistical in nature, but there are other operations more nearly purely logical, and it is the appropriate combinations of these that are difficult to derive. We have tried to make a combination of these two, and the papers listed in the bibliography will help to understand them.

When dealing with objects such as plants, we have a bewildering mixture of quantitative and qualitative types of information. Furthermore our objects (plants) occupy space, and any 2-dimensional representation is bound to introduce an element of artificiality.

Another factor of some importance is that there is no real randomness in our usual classification schemes, though we have to deal with species in some aspects, as randomly distributed. In other words, we have no randomness in our final classification system, but we work from a random type model, hoping to detect patterns by comparisons.

Considerable thought has been given to the process of weighting of characters when they are used by taxonomists. In our system, the individual character has no weight. This will become clearer in next week's talk. Unless some objective evaluation can be demonstrated, it is not wise to assume that one item of information is more important than another. Our statements on weighting at the moment do not have much security--"this is a highly significant Character", "This is fundamental", etc., cannot really be defended by the taxonomist. The idea of non-weighted taxonomy is certainly not new. 200 years ago Michael Adanson made a contribution, and today, a non-weighted system is generally called "Adansonian".

But we do not really have non-weighted systems, in spite of all the talk in this direction. There obviously is some information "more significant" than others--it is a question of evaluation--of testing. Two methods are suggested, and these are (1) deriving a non-weighted classification, and then, after deriving, contrast it with variously weighted ones. This type of work is only possible with a computer. We still have to use intuition at the beginning to know which item(s) of information is considered more valuable--just as in any scientific problem, the individual must know what it is he wants to know--there is no substitution.

(2) A second method of testing the value of a character is by elimination--What happens in a classification by deleting certain items of information. This is a correlary, in essence, to method (1) which doubles, or increases, the weight of a character by some factor.

④ A description of the computer and its hardware is difficult to condense into a short time, but there are certain useful terms that can be explained. Refer to the two books in our office for good, adequate, layman's description of the computer.

1. The method of communicating with the computer takes several forms, the most common being a punched card. Frequently, a modification of the card, for continuing types of data input is a punched paper tape,. Direct communication may be had by use of a typewriter wired directly to the console panel of the computer, where the operator may write instructions to modify or change a program that is in progress. A rapid way to start a computer program going into the machine is by use of a pre-coded electronic tape.

2. Information is coded, so that positive or negative electric impulses (from the hole in the card) impresses itself in one of several types of storage. Each impulse is a "bit". Several bits are needed to handle most information. Most machines are designed to accept one of several numbers of bits, each such number of bits = one complete "word." The 1604 handles 48-bit words. Example--each of our attributes ~~xxxxxx~~ uses 16 bits, and three attributes are stored in a single machine word.

The fastest access storage is called "core" storage. The bit information is stored as "+" or "-", or as "0" or "1", or as "yes" or "no". in doughnut shaped bodies, or on the surfaces of thin plastic films. Info. stored in core is that which is needed for computation purposes, and is needed most rapidly. Access time to core info. is a fraction of a micro-second (less than a millionth of a second). Less rapid access time storage of ~~xxxx~~ less needed info. is either in disc-pac storage (like a phonograph record in some ways) or an electronic tape, the slowest storage method

because of the physical limitations of the tape itself. Even in tape storage, however, time of access is measured in thousandths of a second.

All use of storage units is controlled by a "program" which gives addresses to each piece of information and then directions as to whether the bits are to be stored in core, tape or disc-pacs.

The program is a written statement of the activities in which the computer is directed to move from one activity to the next, in a manner such that additions, subtractions, multiplications, squares, square roots, etc. are done in certain sequences according to a "flow" chart.

To put info. into the machine, one has first to alert the computer to get it ready to accept data and instructions.

The the program is "loaded", and then the data from the cards is placed in appropriate storage. From then on, the activities are controlled by a central processing unit which which places the program instructions, and data in the most efficient storage areas, calls forth each as is needed, instructs and does the computations, tells what the results are, both in a visual display of lights on the control panel and by giving instructions to the printing machine, attached directly to the computer.

The important factors in the computer, for our purposes, are first--accuracy, second--speed, thrid size of storage capacity, fourth--flexibility of operations, and fifth--economy. It sounds #idiculous to talk about economy when we have all heard the sale or rental prices at rates that sound astronomical. What we have to remember is that what we accomplish with a computer just would never get done if we had to hire men to do the calculations.~~thorought~~ You may ask whether it is worth doing, with or w/o a computer. Let us decide this after more discussion.

Programming

Most programming in business procedures requires no more than the establishment of a flow chart, then from this a set of instructions understandable in the computer. Here, we know what the results should be.

In scientific work, however, programming usually requires that a considerable amount of work has to be done before the flow chart is established. We have to work from some model--or pattern--of the thought processes involved in the analysis as they are carried out by the scientist himself. Then there ~~xxxx~~ has to be a statement (or model) which reflects the thinking of the scientist, usually expressed as a series of equations. We do not know what the results should be, and the scientific program must be more flexible than the business program to allow for unexpected results.

Data Recording of Taxonomic or other Information

For the biologist, it is necessary to gather data in any number of places where it would be inconvenient to have the usual equipment for punching cards. We have adapted a gadget designed by IBM, not with biologists in mind, which will permit the biologist to prepare his data for the computer directly, without going through several different data-recording formats. We will describe the method subsequently.

Before actual data recording begins, it is necessary to establish the pattern of, or organization, that will be followed. Our procedures are almost direct take-off from Anderson's methods. Those of you who have read his "Introgressive Hybridization" will be able to follow the procedure easily.

- 1) Assignment of characters and attributes (variables of the character) in advance.
- 2) Establishment of the proper sets of variables (attributes) for each character.
- 3) Assignment of a position to each of the attributes within the characters.
- 4) Division of qualitative variables (attributes by inspection of the range of variation exhibited by the group to be classified.
- 5) Quantitative characters offer less of a problem when assigning attributes.
- 6) Consistent recording of each of the chosen items--a pattern not too frequently

Description of the card.

1. The fields--three, on the left is one field for the attributes, numbering up to 70 attributes that can be recorded on a single card. Frequently, many more attributes will be found than can be punched on just one card.
2. The second field--specimen identification, allowing up to 9999 specimens to be recorded for any one run on the computer.
3. The third field--when more than one card is necessary to record all the attributes of any one specimen. With this card-design, up to 630 attributes for any one specimen can be handled.

Bibliography for Taximetrics Seminar

I. General Computer Information

Berkley, Edmund C. The Computer Revolution. Doubleday, New York, 1962.

International Business Machines Corporation. General Information Manual. Introduction to IBM Data Processing Systems. 1960.

Pfeiffer, John. The Thinking Machine. J. B. Lipincott Company New York. 1962.

II. Taximetrics (Numerical Taxonomy)

Anderson, E. and E. C. Abbe. A quantitative comparison of specific and generic differences in the Betulaceae. Jour. Arnold Arboretum, 15: 43-49. 1934.

Anderson, E. and T. W. Whitaker. Speciation in Uvularia. Jour. Arnold Arboretum, 15: 28-42. 1934.

Chesseman, G. C. and Barridge, N. J. The differentiation of bacterial species by paper chromatography. VII. The use of electronic computation for the objective assessment of chromatographic results. Jour. Appl. Bact. 22(2): 307-316. 1959.

Colwell, R. R., W. J. Wiebe, and J. Liston. A computer study of strain variation within the species Pseudomonas aeruginosa and Pseudomonas fluorescens. Proc. Amer. Assoc. Adv. Sci., Davis, Calif., June, 1961.

Colwell, R. R. and J. Liston. An electronic computer analysis with the electronic computer of some Xanthomonas and Pseudomonas species. Bact. Proc. p. 72. 1961.

Colwell, R. R. and J. Liston. Taxonomic analysis with the electronic computer of some Xanthomonas and Pseudomonas species. Jour. Bact. 62(6): 913-919. 1961.

Gilardi, E., L. R. Hill, M. Turri, and L. G. Silvestri. Quantitative methods in the systematics of Actinomycetales, I. Giorn. Microbiol. 8(4): 203-216. 1960.

Liston, J. Some results of a computer analysis of strains of Pseudomonas and Achromobacter, and other organisms. Jour. Appl. Bact. 23: 391. 1960.

Lysenko, O. Pseudomonas--an attempt at a general classification. Jour. Gen. Microbiol. 25: 379. 1961.

Michener, C. D. and R. R. Sokal. A quantitative approach to a problem of classification. *Evolution* 11(2): 130-162. 1957.

Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Jour. Roy. Statist. Soc. B*, 10: 159-203. 1948.

Rhodes, M. E. The characterization of *Pseudomonas fluorescens* with the aid of an electronic computer. *Jour. Gen. Microbiol.* 25: 331. 1961.

Rogers, David J. Recent endeavors with computers in taxonomy. *The Garden Journal of The New York Botanical Garden*, 11(6): 201-204. 1961.

Rogers, David J. and Taffee T. Tanimoto. A computer program for classifying plants. *Science* 132(3434): 1115-1118. 1960.

Silvestri, L. G., M. Turri, L. R. Hill, and E. Gilardi. A quantitative approach to the systematics of Actinomycetales based on overall similarity. *Symp. Soc. Gen. Microbiol. XII--Microbiol. Classification--pp.* 333-360. 1962.

Sneath, P. H. A. The application of computers to taxonomy. *Jour. Gen. Microbiol.* 17(1): 201-226. 1957.

Sneath, P. H. A. and S. T. Cowan. An electro-taxonomic survey of bacteria. *Jour. Gen. Microbiol.* 19(3): 551-565. 1958.

Sokal, R. R. and C. D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 38(2): 1409-1438. 1958.

Soria, J. V. and C. B. Heiser. A statistical study of relationships of certain species of the *Solanum nigrum* complex. *Econ. Bot.* 15: 245. 1961.

Tanimoto, T. T. and R. G. Loomis. A taxonomy program for the IBM 704. M & A-6; The IBM taxonomy application: I-IV, 1-65. Preliminary write up. 1959; reprinted, 1960.

Van Woerkom, A. J. and K. Brodman. Statistics for a diagnostic model. *Biometrics* 17(2): 299-318. 1961.

III. Systematics (Taxonomy: botanical, zoological, ecological, medical, etc., discussion)

Adanson, M. *Histoire naturelle du Sénégal. Coquillages. Avec la relation abrégée d'un voyage fait en ce pays, pendant les années 1749, 50, 51, 52 & 53.* Bauche, Paris. 1757.

Adanson, M. *Familles des plantes.* 2 vols. Vincent, Paris. 1763.

Anderson, E. *Introgressive Hybridization.* New York. 1949.

- Andrews, C. H. and P. H. A. Sneath. The species concept among viruses. *Nature, London*, 182: 12-14. 1958.
- Beckner, M. The biological way of thought. Columbia University Press, New York, 200 pp. 1959.
- Bigelow, R. S. Monophyletic classification and evolution. *Syst. Zool.*, 5: 145-146. 1956.
- Blackith, R. E. Multivariate statistical methods in human biology. *Sonderdruck, Med. Docum.* 5: 26-28. 1961.
- Boyden, A. Systematic serology: a critical appreciation. *Physiol. Zool.*, 15: 109-145. 1942.
- Boyden, A. Fifty years of systematic serology. *Syst. Zool.*, 2: 19-30. 1953.
- Brodman, K. Diagnostic decisions by machine. I. R. E. Trans. Medical Electronics. ME-7, 216-219. 1960.
- Cain, A. J. Animal Species and Their Evolution. Hutchinson, London, 190 pp. 1954.
- Cain, A. J. The genus in evolutionary taxonomy. *Syst. Zool.*, 5: 97-109. 1956.
- Cain, A. J. Deductive and inductive methods in post-Linnaean taxonomy. *Proc. Linn. Soc. London*, 170 session: 185-217. 1959.
- Cain, A. J. Taxonomic concepts. *Ibis*, 101: 303-318. 1959.
- Cain, A. J. and G. A. Harrison. An analysis of the taxonomist's judgment of affinity. *Proc. Zool. Soc. London*, 131: 85-98. 1958.
- Cain, A. J. and G. A. Harrison. The phenetic similarity of hominoid skulls. *Proc. Soc. Study of Human Biology*, May, 1960.
- Cain, A. J. and G. A. Harrison. Phyletic weighting. *Proc. Zool. Soc. London*, 135: 1-31. 1960.
- Camp, W. H. Biosystematy. *Brittonia*, 7: 113-127. 1951.
- Candolle, A. P. de. Théorie élémentaire de la botanique, ou exposition des principes de la classification naturelle et de l'art de décrire et d'étudier les végétaux. Déterville, Paris, 528 pp. 1813.
- Danzereau, Pierre. A Universal System for Recording Vegetation. Contributions de l'Institut Botanique de l'Université de Montréal. No. 72. 1958.
- Edwards, J. G. A new approach to infraspecific categories. *Syst. Zool.*, 3: 1-20. 1954.

- Ehrlich, P. R. Problems of higher classification. *Syst. Zool.*, 7: 180-184. 1958.
- Ehrlich, Paul R. and Richard W. Holm. Patterns and populations. *Science* 137(3531): 652-657. 1962.
- Gilmour, J. S. L. A taxonomic problem. *Nature*, London, 139: 1040-1042. 1937.
- Gilmour, J. S. L. Taxonomy and philosophy. In J. Huxley (ed.), *The New systematics*. Clarendon Press, Oxford, 590 pp. 1940.
- Gilmour, J. S. L. The development of taxonomic theory since 1851. *Nature*, London, 168: 400-402. 1951.
- Gilmour, J. S. L. and J. W. Gregor. Demes: a suggested new terminology. *Nature*, London, 144: 333. 1939.
- Gilmour, J. S. L. and J. Heslop-Harrison. The deme terminology and the units of micro-evolutionary change. *Genetica*, 27: 147-161. 1954.
- Goodall, D. W. Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Austral. Jour. Bot.*, 1: 39-63. 1953.
- Gregg, J. R. The language of taxonomy. Columbia University Press, New York, 70 pp. 1954.
- Hill, L. R. The Adansonian classification of staphylococci. *Jour. Gen. Microbiol.*, 20: 277-283. 1959.
- James, M. T. An objective aid in determining generic limits. *Syst. Zool.*, 2: 136-137. 1953.
- Kurten, B. A differentiation index, and a new measure of evolutionary rates. *Evolution*, 12: 146-157. 1958.
- Ledley, R. S. and L. B. Lusted. Reasoning foundations of medical diagnoses. *Science* 130: 9-21. 1959.
- Michener, C. D. Some bases for higher categories in classification. *Syst. Zool.*, 6: 160-173. 1957.
- Michener, C. D. and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, 11: 130-162. 1957.
- Olson, E. C. and R. L. Miller. *Morphological Integration*. University of Chicago Press, Chicago, 317 pp. 1958.
- Osborne, D. V. A numerical representation for taxonomic keys. *The New Phytologist* 62(1): 35-43. 1963.

- Overall, J. E. and C. M. Williams. Models for medical diagnosis. Behavioral Sci. 6(2): 134-141. 1961.
- Simpson, G. G. Tempo and Mode in Evolution. Columbia University Press, New York, 237 pp. 1944.
- Simpson, G. G. The principles of classification and a classification of mammals. Bull. Amer. Mus. Natur. Hist., 85: 1-350. 1945.
- Sneath, P. H. A. The construction of taxonomic groups. Symp. Soc. Gen. Microbiol. XII Microbiol. Classification--pp. 289-332. 1962.
- Sokal, R. R. Distance as a measure of taxonomic similarity. System. Zool. 10: 70-79. 1961.
- Sokal, R. R. and F. J. Rohlf. The comparison of dendrograms by objective methods. Taxon 11: 33-40. 1962.
- Sonneborn, T. M. Breeding systems, reproductive methods, and species problems in protozoa. In E. Mayr (ed.), The species problem. Amer. Assoc. Advance. Science, Publ. 50, 395 pp. 1957.
- Van Workom, A. J. Program for a diagnostic model. I. R. E. Trans. Medical Electronics ME-7, 220. 1960.
- Williams, W. T. and J. M. Lambert. Multivariate methods in plant ecology. II The use of an electronic digital computer for association-analysis. Jour. Ecol. 48: 689-710. 1959.

TAXIMETRICS

Seminar, NYBG, April 8, 1963

DEFINITION: Taximetrics

I. PURPOSE OF SEMINAR: to outline our experimental program which is being undertaken to make use of a computer as an aid in storing, sorting and finally presenting taxonomic data in orderly fashion. The question is immediately raised—why a computer as means? Because the machine does mechanically and with fantastic speed what the human being would labor long to accomplish.

Reasons for presenting the exposition at this time:

1. To stimulate thought along taxonomic lines of what classical taxonomic method really is; and to inquire into the value, if any, of our proposed taximetric program. We believe it has merit.
2. To help us by open discussion to formulate the best methods of approach.

II. PROCEDURE OF SEMINAR:

1. Groundwork giving merely theoretical aspects of taxonomy of today with its various ramifications.
2. Brief introduction to the computer program to be used.
3. Trip to view computer and mode of operation.

III. BACKGROUND OF PRESENT-DAY TAXONOMY

1. Definition. Essentially, the T.M. is the process of organizing, ordering and/or arrangement of information about naturally occurring (or cultivated) plants and animals, utilizing largely external traits of the plants or animals.

The purposes of T.M. are various, but in general, attempt to place like plants together, in hierarchies of ascending order, each hierarchy more inclusive than the last.

In other words, classical taxonomy is the assembling of character data, evaluating it, and finally arranging delimiting characters of value for classification of given groups: family, genera, species, etc.

The T.M. deals with the classification aspects of taxonomy. Identification of organisms is one of the major uses of the classification.

The T.M. is basically similar to other correlative endeavors, such as those in ecology, anthropology, sociology, but the biological T.M. is quite distinct in operation, and in other respects. T.M. ^{may and perhaps should encompass} ~~encompasses~~ all that is called "classical", "biosystematics" and "cytotaxonomic" methods.

2. Scope & Procedures.

The T.M. includes some interesting procedures: keys, the species and higher category descriptions in telegraphic style. Condensation of information in the descriptions is packed with information, but requires some skill in interpretation.

The output of the T.M. takes various forms, many incomprehensible to all but the initiated specialists: monographs of varying sizes of taxa; floristic works of differing types (natural geographic areas, or political units); ~~short notes.~~

We are not considering an allied, though important, adjunct to the T.M., the nomenclatorial ^{here} aspects.

3. The value of T.M. is that it allows us to think about the individual or aggregations of individuals (populations), and to provide a framework for various types of endeavors--evolutionary, physiological and ecological, genetical and plant breeding, and many others.

The classical Taxonomic Method is also a means of achieving a better understanding in some instances of these fields.

4. Application of Taxonomic methods (and/or Discussion?)

We will not argue here about the naturalness of, nor the biological validity of, classification. We assume that ^{the} ~~the~~ purposes ^{of classification} are to give us some logical procedure to grasp the ideas of the universe, and for our own special purposes. The arguments concerning the reflection of nature by a classification are so involved in logic, semantics and philosophy that there is little hope of making definite statements here.

The current taxonomic method is perhaps largely responsible for our inability to answer such questions as

~~Such questions as~~ "is a species a naturally-occurring object?", "do genera and families 'exist' in nature?" These may eventually be answered, but the way I see it, until much more data are correlated we cannot even state the questions correctly. ~~The reason I shy away from this type of question are as follows:~~

And the reasons are these:

1. The objects that are the subjects of our classification are not studied in toto. Only certain aspects, called characteristics, which serve our taxonomic purposes are selected from the objects for correlation, in isolation from the object. *these plants* *plant thing*

2. The method of selection of characteristics is largely intuitive.

depending in part on the temperament of the individual, *worker* *Mary*

times we do not select the characteristics most important for classification-- we do not "see" the characters, or for some unstated reason, we reject them as unimportant. When a character is selected, we may then divide it into different states or expressions. There is most frequently no way to evaluate the genetic importance of a selected character. Questions concerning the number of genes involved in a particular character are not often known. Therefore the biological importance is not assessed.

3. Character~~s~~ of taxonomic ~~im~~portance vary from group to group, and we cannot make a one-for-one comparison between the groups. Thus, a species in one genus of one family is defined by morphological characteristics of the flower, in another by the ecological differentiation, and in a third by its incompatibility genetically.

4. Only the broadest type of guide-lines are provided as specified for all organisms, and justifiably so. *these* we may say, after extensive empirical testing, that reproductive structures of plants will provide more reliable characters than will vegetative ones. This does not mean that each classification must consider them in this order.

So much for classical taxonomic methods,

IV. BACK-GROUND FOR "TAXIMETRICS".

Keys have been put into a number of retrieval formats--the most common being the edge-punched Holorith card. A recent attempt (by Osborne) has been made to use numerical notations for keys. In the computer we actually handle key items, but in a much more rapid means than that by cards, ^{Today -} and thus make actually hundreds of comparisons almost simultaneously.

V. CURRENT STATUS & INVESTIGATIONS.

a Our work has been in that section of the taxonomic method beyond the problems of character selection. We assume the validity of the characters, and work on their interrelationship and their value in forming clusters, or groups.

Because we are still in the experimental stage and attempting to arrive at the most valuable and significant characters to select, we have enlisted the aid of a large number of specialists who are concentrating on these interrelationships with respect to the group or groups of plants currently occupying their attention. In one plant group, for example, the use of proportion of calyx-length to corolla-length is being investigated as a possible character; the type of pubescence may prove to be a delimiting factor in another group; etc.

The problems of character-selection are in need of much work, however, and are being studied by a large number of people, ^{over than Botanists.} The behavioral psychologist, ^{is} perhaps at the fore-front of this type of investigation, ^{groups, ourselves included,} ~~as~~ ^{as of possible value in} ~~many others~~ ^{are looking at pattern analysis,} ~~Even the~~ ^{considering} ~~physicists~~ ^{are beginning to have classification problems with their increasing numbers of particles.} ^{Taximetrics.}

VI. PROBLEMS ARISING IN USE OF COMPUTER

In making use of the computer, we are in virgin territory as far as mathematical expressions of the taxonomic method are concerned. The scientific computer models which have been established have been for use in the analytical sciences, and little has been done for the comparative, synthetical sciences.

Before synthetical sciences?

Most of the endeavors and models that are available are in the realm of statistics. (See Sokal, Sokal & Sneath, Rolfe & Sokal)1. Some parts of the taxonomic method [?] are indeed statistical in nature, but there are other operations more nearly purely logical, and it is the appropriate combinations of these that are difficult to derive. We have tried to make a combination of these two, and the papers listed in the bibliography will help to understand them.

When dealing with objects such as plants, unlike [?] we have a bewildering mixture of quantitative and qualitative types of information. Furthermore, our objects (plants) occupy space, and any 2-dimensional representation is bound to introduce an element of artificiality.

Another factor of some importance is that there is no real "randomness" in our usual classification schemes, though we have to deal with species in some aspects, as randomly distributed.

In other words, we have no randomness in our final classification system, but we work from a random type model, hoping to detect patterns by comparison.

Considerable thought has been given to the process of weighting of characters when they are used by taxonomists. *comment right here? on weighting?*
In our system the individual character given to the computer has no weight. This will become clearer in next week's talk. Unless some objective evaluation can be demonstrated, it is not wise to assume that one item of information is more important than another. Our statements on "weighting" at the moment do not have much security----- "this is a highly significant character," "this is fundamental," etc., can not really be defended by the taxonomist.

The idea of a non-weighted taxonomy is certainly not new. 200 years ago Michael Adanson made a contribution *expand this a little - ?*

Thus, today, a non-weighted system is called "Adansonian".

But inspite of all of the talk in this direction, we do not really have ~~/~~ non-weighted systems. There is, obviously, some information "more significant" than others; it is a question of evaluation-- of testing. And here is where the computer comes in.

VILCOMPUTER METHODS OF EVALUATION OF INFORMATION

1. Deriving a non-weighted classification and then, after deriving, contrast it with variously weighted ones. *Example - ?*

understand or call it a weighted system?
This type of work is only possible with a computer. We still have to use intuition, however, at the beginning to know which item (or item) of information is (or are) considered more valuable---- Just as in any scientific problem, the individual must know the problem he is trying to solve; and, within reason, the general area in which the answer is to be found. There is no substitution for this. *Example?*

2. A second method of testing the value of a character is by elimination. That is, what happens to the system of classification if certain items of information are withheld deliberately. This is a corollary, in essence, to method(1) which doubles, or increases, the weight of a character by some factor or other.

VIII. COMPUTER TECHNIQUES.

A description of the computer and its hardware is difficult to condense into a short space of time, but there are certain useful terms that can be explained here (Refer to the two books in our office for good, adequate, [^]lyman's description of the computer.)

1. The method of communicating with the computer takes several forms, the most common being a punched card. Frequently, a modification of the card, for continuing types of data input, is a punched paper tape. Direct communication may be had by use of a type ~~w~~riter, wired directly to the console panel of the computer, where the operator may write instructions to modify or change a program that is in progress. A rapid way to start a computer program going into the machine is by use of a pre-coded electronic tape.

2. Information is coded, so that ^a positive or negative impulse(from the hole in the card) impresses itself in one of several types of storage. Each impulse is a "bit." Several bits are needed to handle most information. Most machines are designed to accept one of several numbers, each such number being equal to one complete ^{machine} "word". The #1604 machine handles 48-bit words. For example, each of our attributes requires 16 bits, and ^{be} thus three attributes ^{be} stored in a single machine word.

3. The fastest access storage is called "core" storage. The bit information is stored as a "+" or "-", or as "0" or "1", or

Digitized by the Hunt Institute for Botanical Documentation

as "yes" or "no", in doughnut-shaped bodies, or on the surfaces of thin, plastic films. Information stored in core is that which is needed for computation purposes, and is needed most rapidly. Access time core information is a ⁿfraction of a micro-second (less than a millionth of a second.) Less rapid access time storage of less frequently needed information is either in "disc-pac" storage (like a phonograph record in some ways) or an electronic tape, the slowest storage method.

because of the physical limitations of the tape itself. Even in tape storage, however, time of access is measured in thousandths of a second.

4. All use of storage units is controlled by a "program" which gives addresses to each piece of information and then directions as to whether the bits are to be stored in core, tape or disc-pacs.

The program is a written statement of the activities in which the computer is directed to move from one activity to the next, in a manner such that additions, subtractions, multiplications, squares, square roots, etc. are done in certain sequences according to a "flow" chart.

To put info. into the machine, one has first to alert the computer to get it ready to accept data and instructions.

The the program is "loaded", and then the data from the cards is placed in appropriate storage. From then on, the activities are controlled by a central processing unit which which places the program instructions, and data in the most efficient storage areas, calls forth each as is needed, instructs and does the computations, tells what the results are, both in a visual display of lights on the control panel and by giving instructions to the printing machine, attached directly to the computer.

IX The important factors in the computer; for our purposes, are first--accuracy, second--speed, third size of storage capacity, fourth--flexibility of operations, and fifth--economy. It sounds ridiculous to talk about economy when we have all heard the sale or rental prices at rates that sound astronomical. What we have to remember is that what we accomplish with a computer just would never get done if we had to hire men to do the calculations. ~~XXXXXXXXXX~~ You may ask whether it is worth doing, with or w/o a computer. Let us decide this after more discussion.

Programming

Most programming in business procedures requires no more than the establishment of a flow chart, then from this a set of instructions understandable in the computer. Here, we know what the results should be.

In scientific work, however, programming usually requires that a considerable amount of work has to be done before the flow chart is established. We have to work from some model--or pattern--of the thought processes involved in the analysis as they are carried out by the scientist himself. Then there ~~must~~ has to be a statement (or model) which reflects the thinking of the scientist, usually expressed as a series of equations. We do not know what the results should be, and the scientific program must be more flexible than the business program to allow for unexpected results.

Data Recording of Taxonomic or other Information

For the biologist, it is necessary to gather data in any number of places where it would be inconvenient to have the usual equipment for punching cards. We have adapted a gadget designed by IBM, not with biologists in mind, which will permit the biologist to prepare his data for the computer directly, without going through several different data-recording formats. We will describe the method subsequently.

Before actual data recording begins, it is necessary to establish the pattern of, or organization, that will be followed. Our procedures are almost direct take-off from Anderson's methods. Those of you who have read his "Introgressive Hybridisation" will be able to follow the procedure easily.

- 1) Assignment of characters and attributes (variables of the character) in advance.
- 2) Establishment of the proper sets of variables (attributes) for each character.
- 3) Assignment of a position to each of the attributes within the characters.
- 4) Division of qualitative variables (attributes by inspection of the range of variation exhibited by the group to be classified.
- 5) Quantitative characters offer less of a problem when assigning attributes.
- 6) Consistent recording of each of the chosen items--a pattern not too frequently followed.

XII Description of the card.

1. The fields--three, on the left is one field for the attributes, numbering up to 70 attributes that can be recorded on a single card. Frequently, many more attributes will be found than can be punched on just one card.
2. The second field--specimen identification, allowing up to 9999 specimens to be recorded for any one run on the computer.
3. The third field--when more than one card is necessary to record all the attributes of any one specimen. With this card-design, up to 630 attributes for any one specimen can be handled.

Perhaps here, it might be fair to stress the fact that programing^m is at this point very cumbersome and time consuming, but, like all automation at the start, the crudities will, with practice and experiment as well as experience, be smoothed out. Eventually, taximetrics should be simplified and in the foreseeable future a truly great contribution to science in areas not even thought of at the present time.

Bibliography for Taximetrics Seminar

I. General Computer Information

Berkeley, Edmund C. The Computer Revolution. Doubleday, New York, 1962.

International Business Machines Corporation. General Information Manual. Introduction to IBM Data Processing Systems. 1960.

Pfeiffer, John. The Thinking Machine. J. B. Lipincott Company New York. 1962.

II. Taximetrics (Numerical Taxonomy)

Anderson, E. and E. C. Abbe. A quantitative comparison of specific and generic differences in the Betulaceae. Jour. Arnold Arboretum, 15: 43-49. 1934.

Anderson, E. and T. W. Whitaker. Speciation in Uvularia. Jour. Arnold Arboretum, 15: 28-42. 1934.

Cheeseman, G. C. and Berridge, N. J. The differentiation of bacterial species by paper chromatography. VII. The use of electronic computation for the objective assessment of chromatographic results. Jour. Appl. Bact. 22(2): 307-316. 1959.

Colwell, R. R., W. J. Wiebe, and J. Liston. A computer study of strain variation within the species Pseudomonas aeruginosa and Pseudomonas fluorescens. Proc. Amer. Assoc. Adv. Sci., Davis, Calif., June, 1961.

Colwell, R. R. and J. Liston. An electronic computer analysis with ~~the electronic computer~~ of some Xanthomonas and Pseudomonas species. Bact. Proc. p. 72. 1961.

Colwell, R. R. and J. Liston. Taxonomic analysis with the electronic computer of some Xanthomonas and Pseudomonas species. Jour. Bact. 82(6): 913-919. 1961.

Gilardi, E., L. R. Hill, M. Turri, and L. G. Silvestri. Quantitative methods in the systematics of Actinomycetales, I. Giorn. Microbiol. 8(4): 203-216. 1960.

Liston, J. Some results of a computer analysis of strains of Pseudomonas and Achromobacter, and other organisms. Jour. Appl. Bact. 23: 391. 1960.

Lysenko, O. Pseudomonas--an attempt at a general classification. Jour. Gen. Microbiol. 25: 379. 1961.

Michener, C. D. and R. R. Sokal. A quantitative approach to a problem of classification. *Evolution* 11(2): 130-162. 1957.

Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Jour. Roy. Statist. Soc. B*, 10: 159-203. 1948.

Rhodes, M. E. The characterization of *Pseudomonas fluorescens* with the aid of an electronic computer. *Jour. Gen. Microbiol.* 25: 331. 1961.

Rogers, David J. Recent endeavors with computers in taxonomy. *The Garden Journal of The New York Botanical Garden*, 11(6): 201-204. 1961.

Rogers, David J. and Taffee T. Tanimoto. A computer program for classifying plants. *Science* 132(3434): 1115-1118. 1960.

Silvestri, L. G., M. Turri, L. R. Hill, and E. Gilardi. A quantitative approach to the systematics of Actinomycetales based on overall similarity. *Symp. Soc. Gen. Microbiol. XII--Microbiol. Classification--pp.* 333-360. 1962.

Sneath, P. H. A. The application of computers to taxonomy. *Jour. Gen. Microbiol.* 17(1): 201-226. 1957.

Sneath, P. H. A. and S. T. Cowan. An electro-taxonomic survey of bacteria. *Jour. Gen. Microbiol.* 19(3): 551-565. 1958.

Sokal, R. R. and C. D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 38(2): 1409-1438. 1958.

Soria, J. V. and C. B. Heiser. A statistical study of relationships of certain species of the *Solanum nigrum* complex. *Econ. Bot.* 15: 245. 1961.

Tanimoto, T. T. and R. G. Loomis. A taxonomy program for the IBM 704. M & A-6; The IBM taxonomy application: I-IV, 1-65. Preliminary write up, 1959; reprinted, 1960.

Van Woerkom, A. J. and K. Brodman. Statistics for a diagnostic model. *Biometrics* 17(2): 299-318. 1961.

III. Systematics (Taxonomy: botanical, zoological, ecological, medical, etc., discussion)

Adanson, M. *Histoire naturelle du Sénégal. Coquillages. Avec la relation abrégée d'un voyage fait en ce pays, pendant les années 1749, 50, 51, 52 & 53.* Bauche, Paris. 1757.

Adanson, M. *Familles des plantes.* 2 vols. Vincent, Paris. 1763.

Anderson, E. *Introgressive Hybridization.* New York. 1949.

- Andrewes, C. H. and P. H. A. Sneath. The species concept among viruses. *Nature*, London, 182: 12-14. 1958.
- Beckner, M. The biological way of thought. Columbia University Press, New York, 200 pp. 1959.
- Bigelow, R. S. Monophyletic classification and evolution. *Syst. Zool.*, 5: 145-146. 1956.
- Blackith, R. E. Multivariate statistical methods in human biology. *Sonderdruck, Med. Docum.* 5: 26-28. 1961.
- Boyden, A. Systematic serology: a critical appreciation. *Physiol. Zool.*, 15: 109-145. 1942.
- Boyden, A. Fifty years of systematic serology. *Syst. Zool.*, 2: 19-30. 1953.
- Brodman, K. Diagnostic decisions by machine. I. R. E. *Trans. Medical Electronics*. ME-7, 216-219. 1960.
- Cain, A. J. *Animal Species and Their Evolution*. Hutchinson, London, 190 pp. 1954.
- Cain, A. J. The genus in evolutionary taxonomy. *Syst. Zool.*, 5: 97-109. 1956.
- Cain, A. J. Deductive and inductive methods in post-Linnaean taxonomy. *Proc. Linn. Soc. London*, 170 session: 185-217. 1959.
- Cain, A. J. Taxonomic concepts. *Ibis*, 101: 303-318. 1959.
- Cain, A. J. and G. A. Harrison. An analysis of the taxonomist's judgment of affinity. *Proc. Zool. Soc. London*, 131: 85-98. 1958.
- Cain, A. J. and G. A. Harrison. The phenetic similarity of hominoid skulls. *Proc. Soc. Study of Human Biology*, May, 1960.
- Cain, A. J. and G. A. Harrison. Phyletic weighting. *Proc. Zool. Soc. London*, 135: 1-31. 1960.
- Camp, W. H. *Biosystematy*. *Brittonia*, 7: 113-127. 1951.
- Candolle, A. P. de. *Théorie élémentaire de la botanique, ou exposition des principes de la classification naturelle et de l'art de décrire et d'étudier les végétaux*. Déterville, Paris, 528 pp. 1813.
- Dansereau, Pierre. A Universal System for Recording Vegetation. *Contributions de l'Institut Botanique de l'Université de Montréal*. No. 72. 1958.
- Edwards, J. G. A new approach to infraspecific categories. *Syst. Zool.*, 3: 1-20. 1954.

- Ehrlich, P. R. Problems of higher classification. *Syst. Zool.*, 7: 180-184. 1958.
- Ehrlich, Paul R. and Richard W. Holm. Patterns and populations. *Science* 137(3531): 652-657. 1962.
- Gilmour, J. S. L. A taxonomic problem. *Nature*, London, 139: 1040-1042. 1937.
- Gilmour, J. S. L. Taxonomy and philosophy. In J. Huxley (ed.), *The New systematics*. Clarendon Press, Oxford, 590 pp. 1940.
- Gilmour, J. S. L. The development of taxonomic theory since 1851. *Nature*, London, 168: 400-402. 1951.
- Gilmour, J. S. L. and J. W. Gregor. Demes: a suggested new terminology. *Nature*, London, 144: 333. 1939.
- Gilmour, J. S. L. and J. Heslop-Harrison. The deme terminology and the units of micro-evolutionary change. *Genetica*, 27: 147-161. 1954.
- Goodall, D. W. Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. *Austral. Jour. Bot.*, 1: 39-63. 1953.
- Gregg, J. R. The language of taxonomy. Columbia University Press, New York, 70 pp. 1954.
- Hill, L. R. The Adansonian classification of staphylococci. *Jour. Gen. Microbiol.*, 20: 277-283. 1959.
- James, M. T. An objective aid in determining generic limits. *Syst. Zool.*, 2: 136-137. 1953.
- Kurten, B. A differentiation index, and a new measure of evolutionary rates. *Evolution*, 12: 146-157. 1958.
- Ledley, R. S. and L. B. Lusted. Reasoning foundations of medical diagnoses. *Science* 130: 9-21. 1959.
- Michener, C. D. Some bases for higher categories in classification. *Syst. Zool.*, 6: 160-173. 1957.
- Michener, C. D. and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, 11: 130-162. 1957.
- Olson, E. C. and R. L. Miller. *Morphological Integration*. University of Chicago Press, Chicago, 317 pp. 1958.
- Osborne, D. V. A numerical representation for taxonomic keys. *The New Phytologist* 62(1): 35-43. 1963.

- Overall, J. E. and C. M. Williams. Models for medical diagnosis. Behavioral Sci. 6(2): 134-141. 1961.
- Simpson, G. G. Tempo and Mode in Evolution. Columbia University Press, New York, 237 pp. 1944.
- Simpson, G. G. The principles of classification and a classification of mammals. Bull. Amer. Mus. Natur. Hist., 85: 1-350. 1945.
- Sneath, P. H. A. The construction of taxonomic groups. Symp. Soc. Gen. Microbiol. XII Microbiol. Classification--pp. 289-332. 1962.
- Sokal, R. R. Distance as a measure of taxonomic similarity. System. Zool. 10: 70-79. 1961.
- Sokal, R. R. and F. J. Rohlf. The comparison of dendrograms by objective methods. Taxon 11: 33-40. 1962.
- Sonneborn, T. M. Breeding systems, reproductive methods, and species problems in protozoa. In E. Mayr (ed.), The species problem. Amer. Assoc. Advance. Science, Publ. 50, 395 pp. 1957.
- Van Workom, A. J. Program for a diagnostic model. I. R. E. Trans. Medical Electronics ME-7, 220. 1960.
- Williams, W. T. and J. M. Lambert. Multivariate methods in plant ecology. II The use of an electronic digital computer for association-analysis. Jour. Ecol. 48: 689-710. 1959.

Colorado State Univ.
Sagina Xi lect.
Apr. 2, 1965

A quantitative approach to taxonomy-

1. An explanation of the title
taxonomy - the science of classification -
in biology (applies to other sciences as well)
quantitative approach - an effort to
discover the logic and rules in taxonomy,
then to ~~use~~ ^{construct} mathematical models of
the rules -
2. Taxonomy - many consider identical with
classification - but the latter is the
process using taxonomic logic.
3. Taxonomy ~~of~~ then treats of the organization
of the real world - into some orderly, "meaningful"
categories -
Meaningful to the biologist implies
an ~~hierarchy~~ ^{hierarchical} system reflecting, as
well as possible, the evolutionary pathways
and evolutionary units.
4. Taxonomy - until quite recently
intuitive - the rules not expressed -
the procedures learned by experience,
seldom taught, except by fiat.

5. Requirements to make a classification useful -
 1. Relates similar objects, and tells by how much these differ from other, less similar objects.
 2. Contains as much information about the related objects as can be reasonably discovered -
 3. Tells how the information about the objects is correlated -
 4. Ranks the objects into categories of ever-increasing size, until all objects are included.
6. Assumptions:
 1. That if one is careful about selecting the defining information concerning the objects, one may use the selected information as defining parameters of the objects.
 2. The more pieces of information we have about each object, the better we may classify the object.
 3. The objects we wish to classify are, in some real way, capable of being related.
 - 4.

7. How do we classify?

1. Take a portion of the universe,
given by some previous classification-
2. Round up the objects representing this
portion - a genus, a family or other group.
3. Select defining attributes of the objects
4. Determine the degrees of similarity
among the objects, expressing the
similarity as some number.
5. Use these similarity numbers
to produce clusters of the objects
6. Arrange the clusters as a series
of nested boxes

8. Some of the problems in performing these
steps-

1. The lack of a suitable, generalized
set of rules.
2. The lack of definitions - what do you
mean when you say "species"?
What is a defining attribute?
3. What is the purpose of classifying?

9. Some methods of attack -

1. Employ a new type of mathematics
developed within the last 100 years
(Boole, 1854, but not discovered till
about 50 years later.)

Colorado State University
INSTITUTE OF ENVIRONMENTAL BIOLOGY

TO: Members
FROM: Executive Committee
SUBJECT: Forthcoming Seminar for the Institute of Environmental
Biology

April 1, 1965
4:00 p.m.

Student Center
Room 207-209

"Computers and Biological Classification"

Dr. David J. Rogers

Dr. David Rogers, Curator of Quantitative Taxonomy, New York Botanical Garden, will speak on the general topic of classification on the basis of many bits of information programmed through computers. His approach has possible application for physiological, ecological, geological, medical, and other data in addition to taxonomic.

This is a substitute program for Dr. Salisbury's talk. Plans for the May meeting will be announced later.

JA/bah

SIGMA XI

April 2

Middle Ballroom
Student Center

Friday Noon Meeting

The Quantitative Approach to Taxonomy

Dr. David J. Rogers

Dr. Rogers is Curator of Quantitative Taxonomy at the New York Botanical Garden. He received his degree from Washington University and is a pioneer in the field of numerical handling of non-numerical data, in particular, "Numerical Taxonomy." His purpose is to use a computer to aid in the correlation of many bits of information, and as a corollary, to provide a classification with the greatest information content. Also, in this sense it can be used to classify ecological, physiological, medical, geological, and other data.

Dr. Richard Ward will introduce our speaker.

April 9

Middle Ballroom
Student Center

Friday Noon Meeting

Pathogenesis of the Respiratory Distress
Syndrome of the New Born

Dr. Fletcher B. Taylor, Jr.

Dr. Taylor is an outstanding scientist noted for his research on enzyme reactions in blood. He received his Ph.D. in physical chemistry. He is presently at the Cardiovascular Research Institute, School of Medicine at The University of California Medical Center but he will join The University of Pennsylvania next fall to head up a group working on blood chemistry. The pathogenesis of a disease which affects premature babies will be described. This disease is characterized by the formation of a hyaline membrane in the alveolae of the lungs which may lead to asphyxia and death.

He will be introduced by Dr. Ralph Baker. We shall have the medical doctors of Fort Collins as guests.

SPECIAL BOTANY SEMINAR

4 P.M., Thursday, April 1

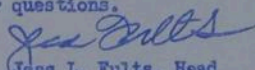
Room 207-209 Student Center

"Computers and Biological Classification"

by Dr. David J. Rogers

Dr. David J. Rogers, Curator of Quantitative Taxonomy, New York Botanical Garden, will speak on the general topic of classification based upon many bits of information programmed through computers. His approach is said to have application for many kinds of data, environmental as well as biological.

Dr. Rogers will also speak to this topic at Sigma Xi the following day, but the longer seminar Thursday afternoon should allow him to expand considerably on his studies and will provide more time for questions.


Jess L. Fults, Head
Botany and Plant Pathology