Clustering of Scientific Journals*

A cluster analysis procedure is described in which 288 journals in the disciplines of physics, chemistry and molecular biology are grouped into clusters. Most of the clusters are easily identified as subdisciplinary subject areas. The data source was the cross citing amongst the journals derived from the *Journal Citation Index (JCI)*, a file derived in turn from the *Science Citation Index (SCI)* ®. The JCI consists of journal by journal tabulation of citings to and from each journal processed in the SCI. Two-step citation maps linking the clusters are presented for each discipline. Within the disciplines the clusters of journals form fully transitive hierarchies with very few relational conflicts.

MARK P. CARPENTER and FRANCIS NARIN

Computer Horizons, Inc.,
1060 Kings Highway North,
Cherry Hill, New Jersey 08034.

• Introduction

This paper describes the partitioning of journals in a scientific discipline into clusters of related journals. The cluster analysis procedure was developed as a tool for the subject classification of journals to a level more precise than that of the discipline.

There are practical uses for an objective journal classification system, both in library science and in the emerging field of science policy studies, where the ability to classify large sets of publications may aid in analysis of scientific capability. There is also an inherent challenge to structuring the literature, since it may well reflect the mosaic of scientific knowledge.

The duality of practical and aesthetic motivation seems to underlie much past research on the scientific literature. This research extends back—at least—to the early parts of this century. One of the first papers based on significant statistical data was Cole and Eales' 1917 analysis of the comparative anatomy literature, as it existed from 1550 to 1860 (*1*). They discussed the rise and fall of publications, the shift of publications from country to country and the interrelationships of publications in science to external economic and political events.

A few years later, in 1923, Hulme published an analysis of the author entries in the *International Catalogue of Scientific Literature* (*2*), and related these to political events in the early decades of this century.

Both the Cole-Eales work and that of Hulme were based on publication counts. The major step of analyzing citations, as opposed to publications, was taken by Gross and Gross in 1927, when they discussed the purchase of journals for a chemical library in terms of a tabulation of citations from the *Journal of the American Chemical Society* (*3*).

Following Gross and Gross's paper there was a burst of papers, often authored by science librarians, attempting to define the importance and dispersions of the various segments of the scientific literature. Many of these papers were reviewed by Stevens (*4*).

After the burst of papers in the 1930's the field became relatively quiescent and stayed so through the war years. Bradford's key 1934 analysis of the importance of a small number of core journals to the search for papers on a specific topic became the basis for much post-war work (*5*).

In the 1950's there was a gradual re-emergence of analysis of the literature and, as science became large in the 1960's, more and more attention was focused on managing the large and rapidly growing scientific enterprise. Price made use of a number of literature counts in devising his macroscopic outlines of the scientific enterprise (*6*). Citation counting began to attract more and more attention as a potential means of structuring the scientific literature. Kessler, in 1964, grouped to-

*The work was supported by the National Science Foundation under Contract NSF-C627. Contributions of Mr. E. Haynes, Evaluation Staff, NSF, as Project Officer and Dr. Bodo Bartocha of NSF in the initiation of the research, are gratefully acknowledged.

gether citations in 20 physics journals, and from this he created a to/from matrix of the percent of cross referencing between the journals (*7*). Van Cott and Zavala attempted to structure the physics literature through a factor analysis based on the subject classification of abstracted articles (*8*). Xhignesse and Osgood studied the citation characteristics of the psychology literature, and created a model showing clusters of highly interrelated journals in psychology (*9*). Papers began to appear suggesting uses of the *Science Citation Index* in studying science, particularly as applied to measuring the importance of individual papers and authors (*10*). We applied many of these techniques to the analysis of the special education literature (*11*).

Central to the present work—clustering scientific journals—is the existence of the *Journal Citation Index* (*JCI*), a special sort of the Institute for Scientific Information's *Science Citation Index* (*12*). The *JCI* contains, in journal by journal lists, a tabulation of all of the citings from each journal (source journal listings) and all of the citings to each cited journal (reference journal listings). The JCI tapes used were for the last quarter of 1969 and contained 729,419 citations from 1821 different journals.

In an earlier paper in this journal (*13*), based on the same JCI data, we showed that two-step maps of the journal literature, in which arrows are drawn from a journal to the journals which it cites first and second-most frequently, allow for the clear delineation of journals to a disciplinary level in biology, biochemistry, chemistry, physics and mathematics, with a relatively small number of linking journals on the boundaries between the disciplines. One of the observations made in the present paper is that the two-step maps, which make use of only a fraction of the citation data for a journal, can be easily compared with the cluster analysis results, that is, that the analytically derived clusters occupy relatively compact areas on the maps.

• The Clustering Process

The process used to divide sets of journals into subject areas has two underlying assumptions: first, that journals which deal with the same subject area will have similar journal referencing patterns; and second, that journals which deal with the same subject area will refer to each other. Using these two assumptions, some relatively straightforward techniques of cluster analysis were applied. Additions and modifications of these methods were made quite empirically. A technique would be tried on a real set of data. Its faults would be deduced from the results it produced. Then changes would be made to the technique which seemed likely to eliminate those faults. In the early stages of this process bad sets of clusters were obvious.

One of the questions plaguing cluster analysis is whether or not a given set of data really contains clusters. Many clustering techniques, including most of those employed here, will produce clusters from any set of data, no matter how dispersed the members. An example of a set of points in a plane may help illustrate the point. A uniform distribution of the points or a randomly generated distribution can be divided into groups by the application of cluster analysis techniques. That these groups are true clusters would be disputed, since the groups would be neither isolated nor compact. On the other hand, members of a group are, on the average, much closer to fellow-members than to non-members, and it is this property of the grouping that is important for the application to scientific journals. What is desired in this case is a grouping of journals which are relatively "near" to each other so that each group represents a subject area. How isolated these groups are from each other, or how compact they are, is relatively unimportant. This is one of the rare cases in which partitioning as opposed to hierarchical classification or clumping (*14*), and dissection as opposed to classification (*15*), are quite satisfactory.

Cluster analysts also worry about the problem of finding the best set of clusters under a certain set of criteria. Since the number of possible partitions increases extremely rapidly as the number of objects increases (for example, over 100,000 possible partitions for a ten object set), it is prohibitive to test all possible partitions. Consequently, many frequently used techniques of cluster analysis are designed only to find a relative maximum of the parameter they try to optimize. That the optimum set of clusters under a given set of criteria may not in fact be reached is of little consequence in the grouping of journals. What is desired is to have the journals divided reasonably into subject areas; the shifting of a few journals, or even the splitting or combining of some clusters, will not significantly alter the conclusions reached in analyzing the relationships of the subject areas. In point of fact, the problem of finding optimal clusters under a given set of criteria seems quite insignificant compared to the problem of choosing a proper set of criteria for whatever data is being analyzed.

The clustering sequence was affected, of course, by limitations in the data. The *Journal Citation Index* consists of counts of references from 1821 citing journals to all the journals they cite, combined and listed by citing journal and by cited journal. A journal cited by another is listed only if it receives five or more citations from that journal. From this data, a journal cross citing matrix was constructed, having as row entries the number of references from a given journal, and as column entries the number of citations to a given journal. The cutoff at five citations eliminated useful data in the case of journals receiving few citations.

The journals representing each discipline to be submitted to cluster analysis had been systematically chosen in the course of other work. The objective was to include the most significant journals in each field, that is,

those which satisfied one of two size criteria: either number of references from, or number of citations to. The search was conducted in several steps with the cutoff point for each criterion lowered in successive steps. Thus in the first pass all journals with $\geq$ 300 citations to them or $\geq$ 500 references from them were incorporated in the list. In the second pass all those with $\geq$ 200 citations to them or $\geq$ 300 references from them were included and so forth. Satisfying only one of the two criteria was sufficient for inclusion. When it appeared that good coverage of a discipline was obtained after a given pass, that discipline was dropped from consideration in the remaining passes. A further requirement, unrelated to the cluster analysis, was that the journal had to have been covered by the *Science Citation Index* from 1965 onward. Table 1 summarizes the results of this process, giving the number of journals chosen and the levels which had to be surpassed for a journal to be included. Complete cluster analysis results for all of these journals in each of these disciplines are available from the authors.

Each discipline, *e.g.*, physics, was treated separately by the cluster analysis. A separate journal cross-citing matrix was constructed for each discipline, so that citing between journals not in the same discipline was not included in the analysis. The number of journals which could be considered at one time was limited by computer capacity.

The three basic tools needed for the clustering process are a measure of similarity between objects being clustered, a measure of the quality of a given set of clusters, and a technique for aggregating objects into clusters which usually utilizes a similarity measure and a cluster quality measure. Many variants of these three types of tools exist. Those used in the clustering process described in this paper will now be discussed. Figure 1 presents an overall picture of the clustering process.

Two measures of the similarity of two journals were incorporated into the final procedures. One was the

TABLE 1. Journal Selection Criteria

Field	No. Journals Chosen	Selection Criteria (in addition to 1965 coverage) Citations To Last 1/4 1969	References From Last 1/4 1969
Molecular Biology	106	$\geq$300	$\geq$500
Chemistry	101	$\geq$200	$\geq$300
Engineering	68	$\geq$200	$\geq$300
Physics	81	$\geq$200	$\geq$300
Systematic Biology	52	$\geq$200	$\geq$100
Mathematics	43	For these last two fields,	
Psychology	43	all journals covered in	
	494	1965 were included.	

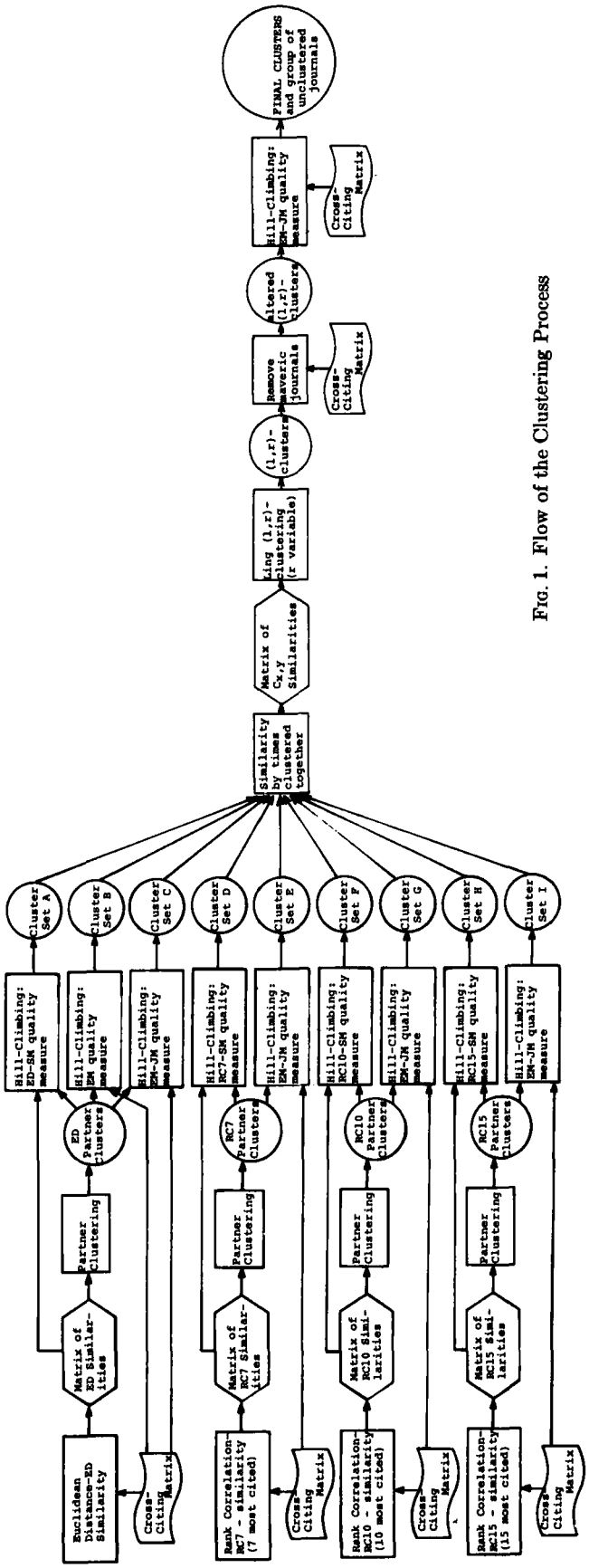

FIG. 1. Flow of the Clustering Process

simple Euclidean distance measure, given by:

$$d_{x,y} = \left[\sum_i (M_{x,i} - M_{y,i})^2\right],$$

where $M_{a,b}$ is an element of the journal cross-citing matrix. It is clear that if $M_{a,b}$ is simply a number of references, the measure $d_{x,y}$ will be useless if the journals x and y are of much different sizes. Hence each row of the matrix M is normalized so that its elements are the percent of the references from the row journal which refer to the column journal.

The second measure of similarity used was taken from Spearman's rank correlation measure:

$$\rho_{x,y} = 1 - \frac{6\left[\sum_i (R_{x,i} - R_{y,i})^2 + T_x + T_y\right]}{N(N^3 - 1)},$$

where $R_{a,b}$ is the rank of journal b in receiving citations from journal a, and T_a is a factor which compensates for tied ranks among the journals referred to by journal a (*16*). The use of a rank correlation measure was suggested by the previous success in obtaining primitive subject groupings of journals through one- and two-step maps.

There are two problems in applying the Spearman measure directly to the data at hand. If the two journals being compared refer at least five times (recall that the JCI cutoff is 5) to different numbers of journals, those journals receiving no references will have a different rank for the one journal than for the other. Secondly, differences in rank become much less significant as the rank increases, since a journal ranked first would typically receive 200 references, the second ranked 150, while the 15th ranked receives 10 and the 25th ranked gets 5. These problems are eliminated by considering only the few journals ranked very highly in being referenced by one of the two journals being compared. There was no clearly optimum number to include, so three rank correlation similarity measures are used, one utilizing the top 7, another the top 10, and the third the top 15 journals cited by either of the two whose similarity is being determined.

The primary technique used in determining clusters is the hill-climbing method. This technique requires a measure of cluster quality, to be discussed below, and, as input, a tentative set of clusters. Under this method each journal is taken in turn, and tried in every existing cluster. If the measure of cluster quality improves by moving the journal to another cluster, the move is made. A hill-climbing "pass" is completed when all journals have been so treated. Such passes continue until one occurs in which no journal is moved.

The set of clusters used to initiate the hill-climbing was produced by a partitioning technique dubbed partner clustering. Each journal is paired with the journal to which it is most similar by a given similarity measure. Clusters are formed so that a journal and its partner always appear in the same cluster but clusters are kept as small as possible given that requirement. Any of the similarity measures may be used. It was found that the initial set of clusters did not have an overpowering effect on the final output of the hill-climbing process, except in terms of the number of clusters produced, since hill-climbing cannot create new clusters and rarely evacuates old ones. Even initial sets created by random assignment produced reasonable results.

Two basic types of cluster quality measure are used. One is simply the average over all journals of the similarity between journals in the same cluster. This type will be referred to as SM. Here, also, any of the similarity measures can be used. The second type is based not on the similarity of the reference structures of clustered journals, but on the amount of cross-citing among clustered journals. For this type of cluster quality measure it is helpful to introduce the concept of the cluster-limited journal cross-citing matrix (CLJCC matrix), which is analogous to the journal cross-citing matrix but limited to the citing among journals within a single cluster. Again, the percentage of references from the row journal to the column journal was used, rather than actual numbers of references, so that size distortions were removed. For each set of clusters there is a corresponding set of CLJCC matrices. A natural measure, which will be called the journal measure, or JM, on which to base hill-climbing decisions would be the average total citing a journal gives to its cluster, not including journal self-referencing. This is simply the average of the sums across the rows of the CLJCC matrices (again, omitting the journal self-referencing, or diagonal elements). Journal self-referencing is eliminated because its size would make it a dominant factor whereas it contains little or no information on subject area. The cluster quality measure JM is unfortunately highly unstable, since journals flock in a chain-reaction manner about highly referenced journals such as the *Physical Review* in physics.

Another natural measure, which avoids this problem, is the entry measure, EM, which simply consists of the average of all the non-diagonal (eliminating self-referencing) entries of the CLJCC matrices. It is thus the average percent of references a journal gives to another member of its cluster. Unfortunately this measure abhors large or small clusters, and will even go so far as to cause the movement of a journal from a large cluster with whose journals it engages in much cross-referencing, to a small one in which it will engage in no cross-referencing whatsoever. The resolution of these problems is to use a combination of the two measures, EM-JM, so that a journal is moved to another cluster if by so doing EM is increased, but JM is not decreased.

Note that all the measures of cluster quality discussed have been broad in scope; they concern averages over all journals rather than just the journal being considered for movement. Rather than moving a journal to the place where it fits best, a journal is moved to the place for which the overall clustering is best.

Many combinations of clustering method and quality measure were tried. Nine of these techniques gave quite reasonable results, each with its relative strengths and weaknesses related to the measures used to produce it. Instead of making an arbitrary choice as to which method and measure was best, a means of combining their results was devised. Hence, all nine techniques are used in the final process; those being clusterings A-I in Figure 1. Three clusterings are produced by partner clustering on the basis of the Euclidean distance similarity measure, then hill-climbing from there, once using Euclidean distance as an SM quality measure, once using EM as the quality measure, and once using EM-JM. The other six clusterings are produced by partner clustering on each of the three altered rank correlation similarity measures and hill-climbing from each of these twice, once using the corresponding rank correlation as an SM quality measure, and the second time using EM-JM as the movement criterion.

These nine clusterings are combined by creating a new similarity measure $C_{x,y}$ which is simply the number of times journal x appears with journal y in the same cluster. Actually, the maximum value of $C_{x,y}$ is 12 rather than 9, since the clusterings, beginning with Euclidean distance partner clustering, are weighted double to give Euclidean distance and rank correlation approximately equal weight. A method due to Ling (*17*) was used to form clusters based on the measure $C_{x,y}$. If a minimum distance value r is specified, clusters are formed so that for every journal, x, in a cluster there is some other member of the cluster, y, for which $C_{x,y} \geq r$, and for any $C_{x,y} \geq r$ the journals x and y are in the same cluster. That is (l,r)-clustering by Ling's terminology. Note that there may be journals that are not in any cluster.

The entire clustering process was programmed for a Honeywell 1648 time-sharing system. At the point in the clustering at which the (l,r)-clustering is done, the operator is able to specify the value of r and observe the result that value produces. He is then free to try another value, until he is able to decide which value is optimal. In trying to maximize the number of clusters while minimizing the number of unclustered journals, the optimal r usually is quite apparent.

Because the characteristic which two journals frequently clustered together have in common may simply be their difference from the rest of the set, rather than a similarity with each other, a few final steps are required to clean the (l,r)-clusters. First, any clustered journal for which the sum of its row and column entries in its CLJCC matrix, excluding self-referencing, is less than two percent is removed to become an unclustered journal. Next, a slightly modified hill-climbing is performed. EM-JM is the movement criterion but of course only clustered journals are considered in determining it. Any clustered journal may be moved to any other cluster, but may not become unclustered. An unclustered journal may move to a cluster if EM-JM is satisfied. As indicated in Figure 1, the output from this step is the final set of clusters.

• Results

The clusters this process has produced are given in Tables 2 to 4 for the fields of physics, chemistry and molecular biology. The journals have been grouped so reasonably that in most cases it is quite easy to attach a subdiscipline label to a cluster based on the journal titles alone. In a few cases the labels are somewhat inadequate, and Tables 2 to 4 should be checked from time to time in subsequent discussion when clusters are referred to by these labels.

Among the physics clusters in Table 2 there is one very large cluster of general physics journals typified by the *Physical Review*. This cluster also encompasses the nuclear physics subdiscipline. A second, small clus-

TABLE 2. Clusters For a Set of 81 Physics Journals

Acoustics
Acustica
J Acoust So
J Sound Vib
Sov Ph Ac R

Minerals
Am J Sci
Am Mineral
Mineral Mag

Geophysics and Space
Ann Geophys
J Atm Ter P
J Geoph Res
Naturwissen
Planet Spac
Pur A Geoph
Rev Geophys
Spac Sci R

General and Nuclear Physics
Am J Phys
Ann Physics
Ann R Nucl
Ark Fysik
Can J Phys
CR Ac Sci B
Helv Phys A
J Math Phys
J Physique
Nucl Phys
Nuov Cim
Phys Lett
Phys Rev
Phys Rev L
Physica
Prog T Phys
Rep Pr Phys
Rev M Phys
Z Phys

German Physics
Ann Physik
Z Ang Phys
Z Naturfo A

Optics
Appl Optics
J Opt Soc

Solid State and Applied Physics
Adv Physics
Appl Phys L
Czec J Phys
I J PA Phys
J Appl Phys
J Phys Ch S
Jap J A Phy
Philos Mag
Phys Fluids
Phys Kond M
Phys St Sol

Geology
Arctic
Geoch Cos A
Geol S Am B

Astronomy and Astrophysics
Astron Astr
Astronom J
Astrophys J
Aust J Phys
B CSAR Belg
Icarus
J Atmos Sci
P Roy Soc A
Sov Astro R

Soviet Physics
DAN USSR
JETP Letter
Opt Spect R
Sov J Nuc R
Sov Ph JE R
Sov Ph SS R
Sov Ph TP R
Sov Ph US R

General Physics
J Phys ABC
J Phys D
J Phys Jap

Fluid Mechanics
J Fluid Mec
Phi T Roy A
Q J R Meteo

Unclustered Journals
J Res NBS A
Nucl Fusion
Rev Ro Phys
Rev Sci Ins
Z Ang Geol

TABLE 3. Clusters For a Set of 101 Chemistry Journals

Chemistry of Solids

Act Chem Sc
Ark Kemi
J Am Ceram
J LessC Met
Suom Kemist

Crystallography

Act Cryst
Soc Ph Cr R
Z Kristall

Metals

Act Metall
ASM T Quart
J I Metals
J Iron St I
J Metals
Mem S A Met
Metall
Russ Met R
T Jap I Met
T Met S AIM
Z Metallkun

Analytical Chemistry

Act Chim H
Analyst
Analyt Chem
Analyt Chim
Atom Ener R
Cereal Chem
Chem Listy
Coll Czech
J Agr Food
JAm Oil Ch
J AOAC
J Chem Educ
J Chromat
J Elec Chem
J Prak Chem
Mikroch Act
Pharmazie
Z Anal Chem

Bulletins & Reviews

Ann Chim Fr
B S Chim Be
B Pol Chim
B S Chim Fr
B S Fr Min
CR Ac Sci C
Q Reviews

Physical Chemistry

Appl Spectr
Ber Bun Ges
J Chem Phys
J Chim Phys
J Mol Spect
J Phys Chem
Molec Phys
Spect Act
T Farad Soc
Theor Chim

General and Organic Chemistry

Angew Chem
Ann Chem
Ann Rp Ch B
B Chem S J
Can J Chem
Chem Ber
Chem Pharm
Chem Rev
Helv Chim A
J Am Chem S
J Chem SABC
J Chem S D
J Hetero Ch
J Org Chem
J Orgmet Ch
Monats Chem
Rec Tr Chim
Tetrahedr L
Tetrahedron

Electrochemistry

Corrosion
J Elchem So

Inorganic Chemistry

Inorg Chem
J Inorg Nuc

Colloids & Polymers

I J Chem
J Coll I Sc
J Pol Sci
Kolloid-Z
Makrom Chem

German Chemistry

Z Anorg A C
Z Chem
Z Naturfo B

Unclustered Journals

Aust J Chem
Colloid J R
Current Sci
Disc Farad
Geoch Int R
Isr J Chem
J Catalysis
J Gen Che R
J Med Chem
J Soil Sci
Magy Kem La
Przemy Chem
Rev Ro Chim
Steroids
T NY Ac Sci
Z Phys Ch F

TABLE 4. Clusters For a Set of 106 Molecular Biology Journals

Physiology and Endocrinology

Act Bio Med
Act Endocr
Am J Physl
Can J Physl
Circul Res
Endocrinol
J Clin Inv
J Endocr
P Soc Exp M
Pharm Rev
Physiol Rev

Microbiology

Ann R Micro
Bact Rev
Can J Micro
J Bact
J Gen Micro
Path Biol

Biochemistry

Analyt Bioc
Ann In Past
Arch Bioch
B S Chim Bi
Bioc Biop A
Bioc Biop R
Biochem
Biochem J
Can J Bioch
CR Ac Sci D
CR Soc Biol
Eur J Bioch
H-S Z Physl
J Biochem
J Biol Chem
Post Bioch

Plant Biology

Aust J Biol
J Exp Bot
Plant Cel P
Plant Physl

Basic Molecular Biology

Act Med Oka
Agr Biol Ch
Arch G Vir
Arch Mikrob
B Ital Biol
Bioch Pharm
Biophys J
Biopolymers
Cold S Harb
Comp Bioch
Exp Cell Re
Experientia
Fed Proc
J Cell Phys
J Gen Physl
J Mol Biol
J Sci Ind R
J Theor Bio
Nature
P NAS US
P Roy Soc B
Science
Virology

Cancer Research

Br J Canc
Cancer Res
J Nat Canc
Neoplasma

Nutrition

Br J Nutr
J Nutr
P Nutr Soc

Pharmacology

Ann Med Exp
Arch I Phar
Br J Pharm
J Pharm Exp
J Pharm Pha
J Pharm Sci
Jap J Pharm
N-S Archiv

Genetics

Chromosoma
Genet Res
Genetics
Hereditas
Heredity

Clinical Chemistry

Clin Chem
Clin Chim A

Histology and Cytology

Exp Mol Pat
Histochemie
J Cell Biol
J Hist Cyto
J Microscop
J Ultra Res
Protoplasma

Radiation Research

Int J Rad B
Radiat Res

Unclustered Journals

Ann NY Acad
Ann R Bioch
Arznei-For
Exp Neurol
I J Ex Biol
Ital J Bioc
J Anat
J Comp Neur
J Exp Zool
J Lipid Res
J Neurochem
Lloydia
Qual Plant
Sci Am
Z Ernahrung

ter, centered on the primary British physics journal, *Journal of the Physical Society,* must also be called a general physics cluster. There is a good division of journals in three closely related fields into three clusters: astronomy and astrophysics, geophysics and space, and minerals. A fourth group appears by title to be geological in nature. There are small but clear groups for optics and acoustics, while there is a large and rather varied group of solid state and applied physics journals. There is a small cluster which appears to be fluid mechanics.

Clusters formed on the basis of referencing structure need not necessarily reflect only subdisciplines. Clusters characterized by nationality can be formed, as can be seen by the strong Soviet physics group, and a second group whose major characteristic is its Germanic origin.

The chemistry clusters listed in Table 3 also include one very large cluster of journals basic to the field (as in physics typified by a large American journal), in this case the *Journal of the American Chemical Society,* and also as in physics, encompassing a large subdiscipline, this time organic chemistry. There is also a large, strong

physical chemistry cluster and a large analytical chemistry group which includes journals from subdisciplines on the periphery of chemistry which largely rely on analytical methods, such as the *Journal of Agricultural and Food Chemistry* and the *Journal of the American Oil Chemists Society*.

The metallurgy cluster is very distinct. For subdisciplines related to metallurgy, there is a strong crystallography cluster, and also distinct clusters for the chemistry of solids and for electrochemistry. There are also small but strong clusters for inorganic chemistry and for colloids and polymers. As in physics there is a cluster of national character, that being German. The Soviet journals do not form their own cluster in chemistry, probably because a number of the important Soviet chemistry journals were not in the sample since they were not covered by ISI in 1965. In the chemistry set of clusters there appears a new type of cluster, one based not on subdicipline or national characteristics, but on a characteristic type of article. This is the cluster of journals consisting of reviews and short articles, though there is also a strong French character to the cluster. The group of unclustered journals contains three classes of journals: journals from subdisciplines on the periphery of chemistry, foreign journals, and multi-disciplinary journals in which chemistry articles, although probably being a plurality of all articles, are still a minority. Any of these three characteristics can make a journal's referencing structure somewhat unique to the set.

There are quite a few unclustered journals in these fields. Some unclustered journals had so few references in the last quarter of 1969 that there is little information about their referencing structure available to the programs.

In molecular biology, Table 4, there is, again, a large cluster representing general molecular biology. Although somewhat varied in scope, it is centered around such journals as *Nature, Proceedings of the National Academy of Sciences USA,* and *Science*. This cluster is quite distinct from the very large biochemistry cluster. There are separate clusters for the related subdisciplines of microbiology and histology—cytology. A strong pharmacology cluster is separate from a large physiology and endocrinology cluster, while nutrition forms a small cluster of its own. Plant biology, genetics, cancer research, radiation research, and clinical chemistry all form small separate clusters. Comments at the end of the preceding paragraph on the unclustered chemistry journals could be applied equally well to the group of unclustered molecular biology journals.

Cluster cross-citing matrices analogous to the journal cross-citing matrix are presented in Tables 5 to 7 for the three fields. Each element is the number of citations given to journals in the column cluster by journals in the row cluster. Because the computer output for the cluster cross-citing was in the form of percent of references-from, the numbers in Tables 5 to 7 are accurate only to within 0.05 percent of the total number of references-from as given in the final column of the tables.

One of the tools used in our previous paper (*13*) on the relationships among journals is the two-step map. In the two-step map arrows are drawn from each journal to the two journals, other than itself, to which it gives the most references. A considerable amount of organization of the journals is required to make the maps while minimizing arrow crossings and arrow length. These maps have been drawn for the journal sets for physics, chemistry, and molecular biology, and they

TABLE 5. Cluster Cross-Citing Matrix—Physics

From \ To	Acoustics	Minerals	Geoph-Space	Genl Nuc Phys	German Phys	Optics	Sol St-Appl	Geology	Astrn. Astrp	Soviet Phys	Genl Phys	Fluid Mech	Unclust J's	Total References From
Acoustics	833	—	—	12	—	—	41	—	24	5	—	22	—	2430
Minerals	—	262	25	—	—	—	—	39	—	6	—	—	—	1482
Geoph-Space	9	17	2144	300	17	—	94	60	429	34	17	77	9	8575
Genl Nuc Phys	—	—	96	27404	193	96	1399	—	531	1254	772	—	96	48247
German Phys	—	—	5	404	357	—	125	—	44	18	26	—	5	2609
Optics	—	—	13	155	8	703	68	—	71	13	16	—	26	2624
Sol St-Appl	18	—	18	3269	146	110	3233	—	274	603	457	55	164	18263
Geology	13	109	138	—	—	—	—	401	—	—	—	—	—	3211
Astrn. Astrp	—	—	159	408	10	60	70	10	3436	80	30	70	10	9960
Soviet Phys	—	45	22	2739	22	45	674	—	112	7498	112	—	—	22450
Genl Phys	—	—	—	1338	19	24	402	—	160	87	722	19	—	4849
Fluid Mech	6	—	14	6	—	—	80	—	104	—	—	321	—	1958
Unclus J's	—	—	—	191	—	—	58	—	7	11	7	—	226	1868

TABLE 6. Cluster Cross-Citing Matrix—Chemistry

From \ To	Chem Solids	Crystallog	Metals	Analyt Chem	Bulls & Revs	Phys Chem	Gen & Org Chem	Electrochem	Inorg Chem	Coll-Polym	German Chem	Unclust J's	Total References From
Chem Solids	991	200	63	54	34	342	850	15	68	10	63	29	4884
Crystallog	33	1011	5	—	—	84	183	—	33	—	—	—	2539
Metals	52	20	2026	13	7	39	20	52	—	—	7	—	6536
Analyt Chem	115	16	—	3612	49	492	1592	49	164	16	66	66	16416
Bulls & Revs	64	74	21	96	2116	819	3339	11	128	—	32	64	10634
Phys Chem	74	148	—	197	222	9800	2635	74	172	74	222	443	24622
Gen & Org Chem	397	529	—	529	859	3305	38473	66	1058	66	331	463	56104
Electrochem	18	—	14	27	—	127	64	337	5	—	5	32	1767
Inorg Chem	52	161	—	62	41	451	1437	—	1063	—	104	47	5187
Coll-Polym	25	17	—	25	34	729	1314	25	17	1890	8	34	8475
German Chem	27	95	—	27	37	203	664	3	41	—	441	41	3389
Unclust J's	20	29	—	39	29	549	1450	—	98	10	20	1087	9796

show that the organization in the models and in the cluster analysis are in strong agreement. There is, of course, a great deal of flexibility in the positioning of journals in the two-step maps. While it is natural that the clusters, based on referencing structure, should be strongly influenced by which journals its members cite first and second most highly, the ease with which clusters may be outlined on suitably constructed two-steps maps indicates the presence of a surprisingly consistent structure within each field.

The two-step map technique may also be used to illustrate the relationships among the clusters. Figure 2 contains two-step cluster maps for the three fields, with solid arrows pointing to the most highly cited cluster, dotted arrows pointing to the second most highly cited. Chemistry and molecular biology are strongly bipolar, while physics has something of a third pole where clusters near astrophysics and geophysics relate highly to each other. In chemistry, the general and organic chemistry cluster is the dominant pole since all but one of its arrows are due to being cited most highly whereas all but one of the other pole's arrows are due to being cited second most highly. Similarly, the dominant poles in physics and molecular biology are general and nuclear physics, and basic molecular biology.

The six clusters receiving the most arrows in Figure 2

TABLE 7. Cluster Cross-Citing Matrix—Molecular Biology

From \ To	Physl-Endoc	Biochem	Microbiol	Plant Biol	Bas Mol Bio	Cancer Res	Nutrition	Pharmacol	Genetics	Clin Chem	Hist-Cytol	Radiat Res	Unclust J's	Total References From
Physl-Endoc	4455	1827	60	—	1867	141	80	401	—	60	100	20	301	20072
Biochem	826	21227	534	147	6606	146	49	97	—	97	437	—	923	48575
Microbiol	114	1835	2404	23	1983	34	—	—	205	—	103	—	171	11395
Plant Biol	—	568	—	657	356	—	14	—	44	—	14	—	10	3421
Bas Mol Bio	827	4529	473	79	10633	315	39	276	197	—	433	39	394	39381
Cancer Res	163	262	9	—	511	954	9	—	—	—	113	14	59	4522
Nutrition	138	341	—	—	118	4	350	—	—	—	7	—	20	2227
Pharmacol	634	364	—	—	662	19	—	2295	—	9	—	—	177	9330
Genetics	—	139	135	—	577	—	—	—	876	—	36	4	15	3649
Clin Chem	62	265	—	—	69	—	—	5	—	212	—	—	11	1557
Hist-Cytol	104	476	7	13	606	59	—	—	26	—	1389	—	72	6520
Radiat Res	38	127	8	—	287	34	—	—	20	—	12	365	20	1982
Unclust J's	476	1505	38	19	1333	95	38	76	—	38	171	—	1238	19048

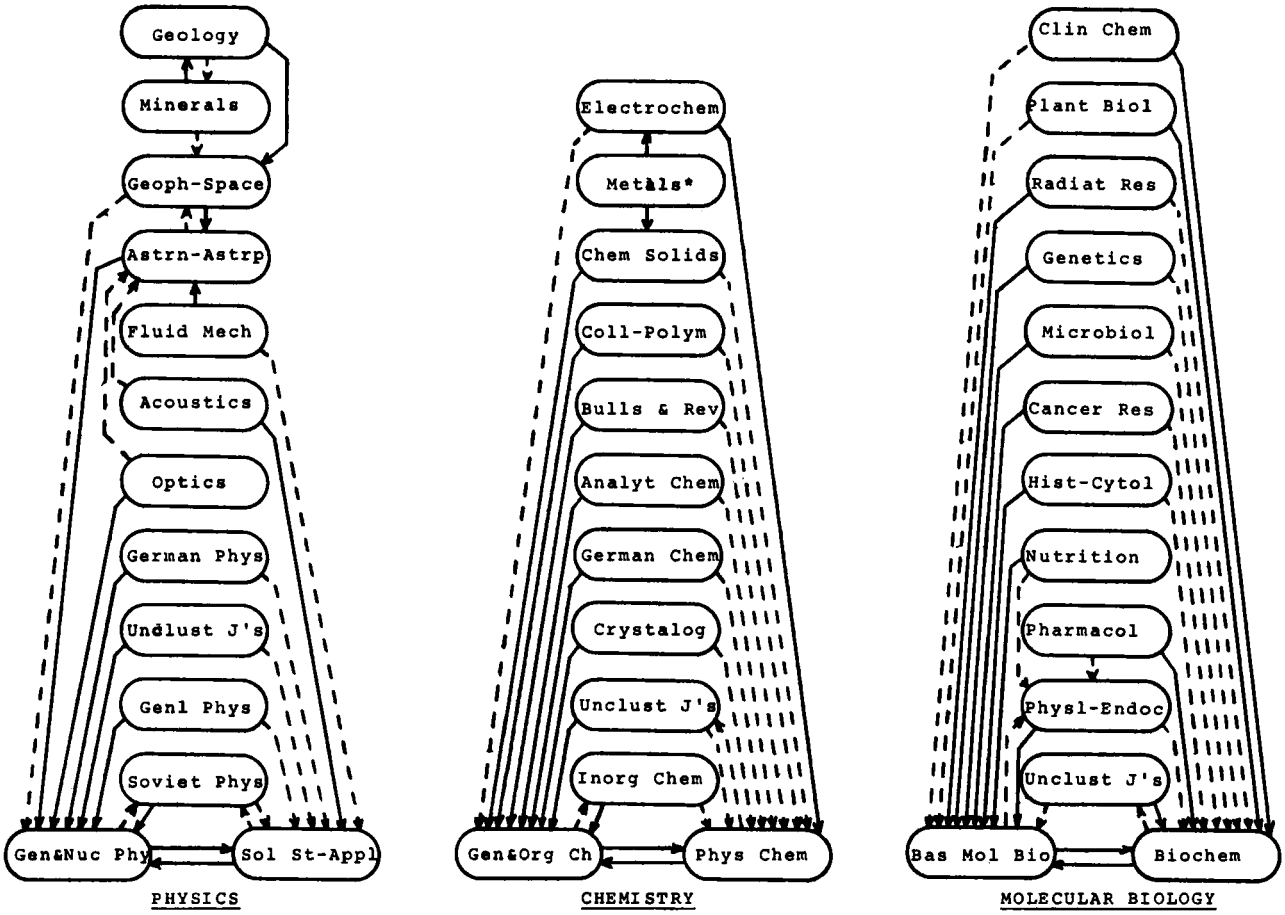

FIG. 2. Two-Step Cluster Citation Maps

are all large clusters, that is, much material is published in the journals which they contain. One indirect measure of the size of a journal or cluster is the number of references from the journal, or the journals in the cluster, since the average number of references per paper is reasonably constant within a discipline. This number of references-from includes self-referencing. Each of the six polar clusters in Figure 2 has first or second most references-from in its field and in addition in physics and chemistry the dominant poles have the most references-from in their fields. Although it is instructive to know which clusters a given cluster refers to most often, it would also be informative to know which clusters it refers to most on a relative basis, with size discounted. One can construct a relative cross-citing measure by dividing the percent of the citations from Cluster A which refer to Cluster B by the percent of the material in the discipline which appears in Cluster B, as measured by the number of references-from. A new two-step map can be constructed using this relative citation measure, as in Figure 3. The structure is now much more complex. The relatively heavy dependence of a subdiscipline on related subject areas is now clear. In molecular biology the structure is particularly clear. Many of the dependencies are both reasonable and interesting. Basic molecular biology, and biochemistry, are still the most highly cited clusters. Some other relationships are: radiation research depends most highly on cancer research, but the reverse is not true; cancer research and histology-cytology are mutually dependent; plant biology depends on genetics while microbiology and genetics are mutually dependent; pharmacology, clinical chemistry, and nutrition all depend on physiology-endocrinology.

Another method of showing citation relationships among journals, which may be used similarly for clusters, is the hierarchy. In constructing a cluster percent citation hierarchy, cluster x is placed higher in the hierarchy than cluster y if x gives a greater percentage of its references to y than y does to x. Thus in the hierarchies in Figure 4 the clusters frequently referred to appear at the bottom. If at least one of a pair of clusters gives the other 0.1 percent or more of its references a hierarchical relationship is said to exist between them. There are $n(n\text{-}1)/2$ distinct pairs among n clusters. Most of these pairs will have a hierarchical relationship.

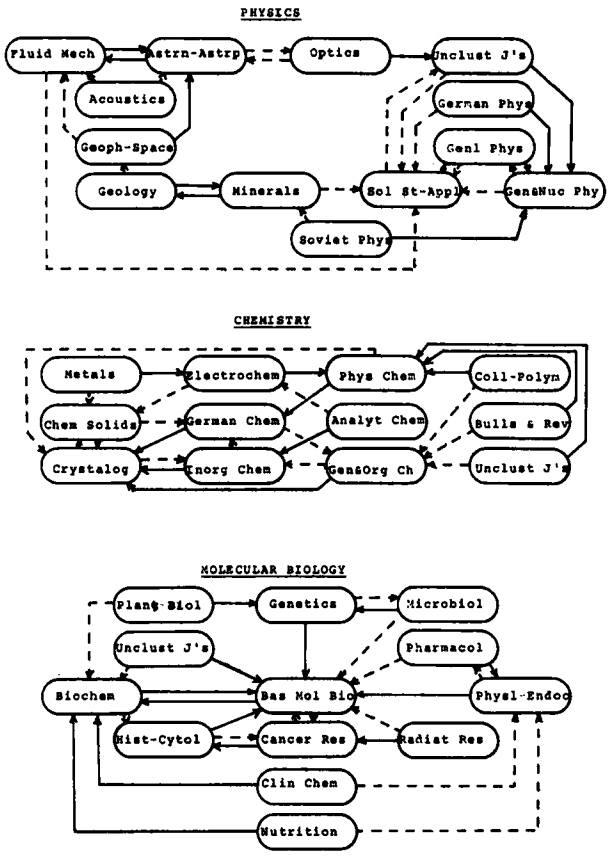

FIG. 3. Two-Step Cluster Relative Citation Maps

TABLE 8. Data on Cluster Percent Citation Hierarchies

	Physics	Chemistry	Molecular Biology
(a) Cluster pairs	78	66	78
(b) Hierarchial Relationships (>.1%)	53	62	57
(c) (b) at percent of (a)	67.9%	93.9%	73.1%
(d) Significant Hierarchical Relationships (>1%)	29	35	42
(e) (d) as percent of (a)	37.2%	53.0%	53.8%
(f) Conflicts	1	3	1
(g) (f) as percent of (b)	1.9%	4.8%	1.7%
(h) Significant Conflicts	0	0	0

Conflicts

Cluster A	Should be	Cluster B	Citations from A to B	Citations from B to A
Physics				
Astrn-Astrp	Above	Geology	0.1%	0.0%
Chemistry				
Electrochem	Below	Bulls & Revs	0.0%	0.1%
German Chem	Below	Metals	0.0%	0.1%
Unclust J's	Below	German Chem	0.2%	0.4%
Molecular Biology				
Plant Biol	Below	Microbiol	0.0%	0.2%

If at least one of a pair refers to the other at least 1 percent of the time the relationship is said to be significant. In constructing a hierarchy, it is of course not always possible to order the clusters in such a way that all relationships are correctly implied. For instance, if cluster x appears higher in the hierarchy than cluster y, and x gives y A percent of its references, while y gives x B percent of its references, and $B\text{-}A > 0$ percent, then a conflict is said to exist. If $B\text{-}A > 1$ percent, the conflict is called significant. Data concerning these concepts is presented in Table 8. The number of relationships and significant relationships is quite high, meaning that the hierarchies are highly specified. Yet there are very few conflicts, and all of these are at a level of less than 1 percent.

It should be kept in mind that a hierarchy constructed in this manner is influenced by the size of the clusters. In fact, the Kendall rank correlation measure τ (*18*) for rank in the hierarchy (bottom=1) versus rank in size is 0.46, 0.48, and 0.63 for the physics, chemistry, and molecular biology hierarchies of Figure 4, corresponding to probabilities of occurrence by chance on the order of 0.01 for lists of that length. This size factor is not present if actual numbers of citations are used rather than percents of citations. Hierarchies based on actual numbers can be constructed in the same

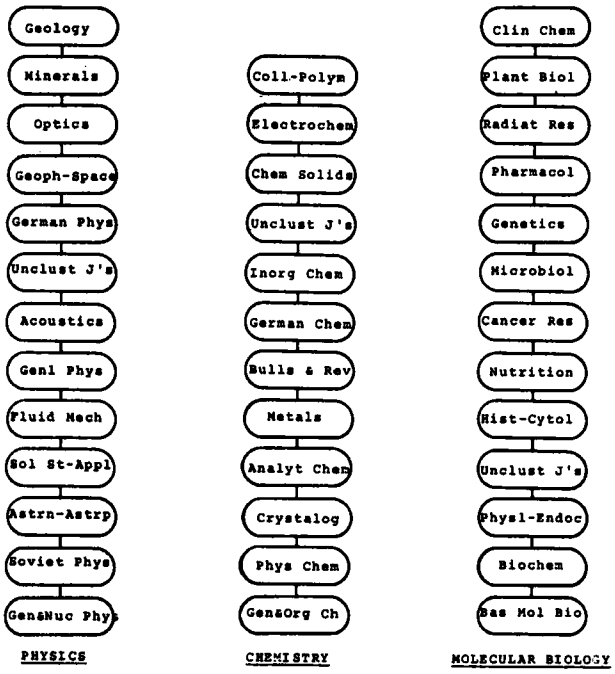

FIG. 4. Cluster Percent Citation Hierarchies

manner as described above for those based on percent of references. Figure 5 presents these citation hierarchies. A pair will have a relationship in the citation hierarchy if and only if it did in the percent citation hierarchy. For a relationship between journals x and y significance is now defined to occur if $A+B \geq 1/10\ s/n^2$ where A is the number of citations from x to y, B is the number from y to x, s is the total number of references from all clusters to all clusters in the field, and n is the number of clusters. Thus s/n^2 is the mean value of an element in the cluster cross-citing matrix. The choice of the fraction (1/10) in this expression was arbitrary, but was chosen so that the number of significant relationships would be approximately the same as it was for the percent citation hierarchy. If there is a conflict between journals x and y, it is considered significant if $|A-B| \geq 1/10\ s/n^2$. Table 9 contains data on the hierarchies of Figure 5.

As in the case of the two-step maps, removing the size effects removes some of the order among the subdisciplines, as a few more conflicts appear. But the hierarchic order is still very strong and it is clear that the conflicts, as listed in Table 9, are far from significant. Lack of conflicts indicates that in molecular biology the clusters relate to each other in the most well-structured manner. That the artificial size effects have been eliminated is indicated by the fact that the Kendall rank correlations of size versus hierarchic rank are now 0.13, 0.09, and 0.31. The order of clusters undergoes a great deal of shifting when size effects are removed, but few clusters undergo extremely large changes in hierarchic rank.

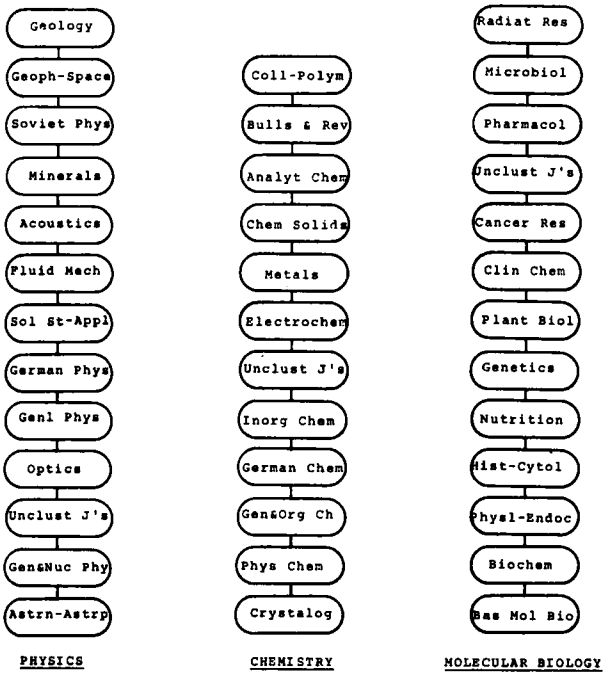

FIG. 5. Cluster Citation Hierarchies

TABLE 9. Data on Cluster Citation Hierarchies

	Physics	Chemistry	Molecular Biology
(a) Cluster pairs	78	66	78
(b) Hierarchical Relationships	53	62	57
(c) (b) as percent of (a)	67.9%	93.9%	73.1%
(d) Significant Relationships	34	39	41
(e) (d) as percent of (a)	43.6%	59.1%	52.6%
(f) Conflicts	9	6	1
(g) (f) as percent of (b)	17.0%	9.7%	1.7%
Significant Conflicts	0	0	0

Conflicts

Cluster A	Should be	Cluster B	Citations from A to B	Citations from B to A
Physics ($s/n^2 = 395$)				
Geoph-Space	Below	Minerals	17	25
Optics	above	Geoph-Space	13	0
Optics	above	German Phys	8	0
Astrn-Astrp	above	Geology	10	0
Soviet Phys	below	Acoustics	0	5
Fluid Mech	below	Genl Phys	0	19
Unclust J's	below	Astrn-Astrp	7	10
Unclust J's	above	Soviet Phys	11	0
Unclust J's	above	Genl Phys	7	0
Chemistry ($s/n^2 = 642$)				
Analyt Chem	below	Metals	0	13
Electrochem	above	Chem Solids	18	15
Electrochem	below	Gen & Org Ch	64	66
German Chem	above	Bulls & Rev	37	32
German Chem	below	Phys Chem	203	222
Molecular Biology ($s/n^2 = 484$)				
Unclust J's	below	Genetics	0	15

• CONCLUSIONS

A cluster analysis procedure has been described in which 288 large and highly cited journals in physics, chemistry and molecular biology have been grouped into clusters of related journals. Many of these clusters can be identified as representing subdisciplinary subject areas although a few clusters seem to have more of a national, than a subject character. Most of the clusters occupy relatively compact areas on suitably constructed two-step journal maps. The interrelationships of the clusters themselves show a large degree of structure, through cluster maps and cluster hierarchies which are quite analogous to journal maps and journal hierarchies. The key clusters in chemistry are a general and organic cluster, and a physical chemistry cluster. In molecular biology the key clusters are a general molecular biology cluster, and a biochemistry cluster. For physics the key clusters are a general and

nuclear physics cluster, and a solid state and applied physics cluster. An astronomy and astrophysics cluster also appears as a third nucleus in physics. There are strong hierarchical relationships amongst the clusters, whether an emphasis on cluster size is included or not, with a relatively small number of conflicts in hierarchical position. At the cluster level the citations in the scientific literature seem to exhibit the same high degree of organization found in dealing with individual journals.

References

1. COLE, F. J. and N. B. EALES, "The History of Comparative Anatomy," *Science Progress,* 11: 578–596 (1917).
2. HULME, E. W., *Statistical Bibliography in Relation to the Growth of Modern Civilization,"* London, Grafton, (1923).
3. GROSS, P. L. K. and E. M. GROSS, "College Libraries and Chemical Education," *Science,* 66: 385–389 (1927).
4. STEVENS, R. E., "Characteristics of Subject Literatures," Assoc. of College and Reference Libraries Monograph No. 6: 10–21 (1953).
5. BRADFORD, S. C., "Sources of Information on Specific Subjects," *Engineering* (London), 137: 85–86 (1934).
6. PRICE, DEREK J. DE SOLLA, *Little Science Big Science,* New York, Columbia University Press, (1963). *Science Since Babylon,* New Haven, Yale University Press, (1961).
7. KESSLER, M. M., "Some Statistical Properties of Citations in the Literature of Physics," M. E. Stevens, Ed., *Statistical Association Methods in Mechanized Documentation* (Symposium Proceedings, 1964), Washington Nat. Bur. Std. 193–198 (1965).
8. VAN COTT, H. P. and A. ZAVALA, *Extracting the Basic Structure of Scientific Literature,* American Institute for Research (1967).
9. XHIGNESSE, L. V. and C. E. OSGOOD, "Bibliographical Citation Characteristics of the Psychological Journal Network in 1950 and in 1960," *American Psychologist,* 22: 778–791 (1967).
10. GARFIELD, E., "Citation Indexing for Studying Science," *Nature,* 227: 669–691 (1970) and "Citation and Distinction," 242: 485 (1973).
11. NARIN, F. and D. GARSIDE, "Journal Relationships in Special Education," *Exceptional Children,* 695–704 (1972).
12. The *Journal Citation Index* is a series of statistical tables derived from the *Science Citation Index. JCI* is available from the Institute for Scientific Information, Philadelphia. The *JCI* has been described recently by E. GARFIELD in "Citation Analysis as a Tool in Journal Evaluation," *Science* 178: 471–479 (1972).
13. NARIN, F., M. CARPENTER and N. C. BERLT, "Interrelationships of Scientific Journals," *Journal of the American Society for Information Science,* 23: 323–331 (1972).
14. CORMACK, R. M., "A Review of Classification," *J. Royal Statistical Society,* Series A, 134: 321–353 (1971).
15. KENDALL, M. G., "Discrimination and Classification," *Proc. Symp, Multiv. Analysis,* Dayton, Ohio, 165–185 (1966).
16. KENDALL, M. G., *Rank Correlation Methods,* London: Griffiin, 38, (1968).
17. LING, R. F., *Cluster Analysis,* Yale University Department of Statistics, Technical Report No. 18 (Jan. 1971).
18. KENDALL, M. G., *op. cit.,* 5, 35, 51.