Scientometrics, Vol. 2, No. 4 (1980) 277–301

CITATION CONTEXT ANALYSIS OF A CO-CITATION CLUSTER: RECOMBINANT-DNA

H. SMALL, E. GREENLEE

Institute for Scientific Information, 3501 Market Street, University City Science Center, Philadelphia, Pa. 19104 (USA)

(Received October 9, 1979)

The techniques of co-citation clustering and citation context analysis are combined to concretely define the shared knowledge within a research specialty. The cluster for a large and fast moving biomedical specialty, recombinant-DNA, is presented in terms of the highly cited documents comprising it and their co-citation links. By examining citation contexts in the papers citing the highly cited documents, it is possible to label each of the documents in the cluster with its specific cognitive meaning for the citing authors. Co-citation contexts are used to reveal the relationships among the concepts symbolized by the highly cited documents, providing a cognitive equivalent of the co-citation links. This may open a new way to the investigation of the logic of conceptual change at the specialty level.

Introduction

If we take seriously a number of recent sociological descriptions of science as a group endeavor in which knowledge evolves by a process of negotiation and consensus formation,[1] then a key remaining task is to devise methods for making manifest exactly what knowledge is negotiated and shared. We are beyond the point where we can simply assert that a "paradigm" exists in this or that field. If progress is to continue in the study of scientific specialties, systematic methods must be devised for delineating exactly what knowledge or cognitive elements are involved in a given case.

The study of scientific specialties has recently seen the introduction of a number of quantitative techniques, such as cluster analysis,[2] multidimensional scaling,[3] factor analysis,[4] and block modelling.[5] Many studies have utilized large computerized files of citation data as input to the statistical techniques. The combination of large files, automation, and sophisticated manipulative procedures has made it possible for the analyst to pose new questions concerning both the micro- and macro-structure of science. For example, we may now ask: What are the specialties of science, and where should one reasonably draw boundaries? Also: How are specialties of science, interrelated? The new quantitative approach has allowed us to derive measures of the social and cognitive structure of science. Such measures can

then be compared with the more traditional approaches to the study of science involving historical/narrative accounts, participant accounts, or data from questionnaires.

While it was clear in earlier work on co-citation, that changes in certain statistical and structural characteristics of co-citation clusters could be interpreted to indicate events in the history of the specialty, such as conceptual shifts, simultaneous discoveries and sub-specialization, we lacked a method for defining in a very precise way the cognitive significance of clusters and their constituent documents and linkages. A solution to this problem was suggested by a number of studies which were beginning to examine citation passages (the language in which references are embedded) in an attempt to classify references according to their role or function in scientific papers.[6-9]

Most of these analyses have attempted to classify references, using a variety of categories, according to function or motivation for citing. Another approach is to use the citation context as a way of studying the meanings or interpretations of the cited works, i.e., the concepts associated with the references.[10] The latter approach will be taken in this paper, since our chief concern will be in laying out the cognitive structure of a research area.

The present study will apply these techniques to one of the most controversial and fast-moving research areas of the recent period, recombinant-DNA. We will attempt to join for the first time techniques of co-citation cluster analysis and the citation context analysis to yield a detailed conceptual map of the area. One of our aims is to see if the technique of citation context analysis can be extended to co-citation contexts, and thereby open the way to a systematic analysis of the logical structure of shared knowledge in scientific specialties.

Clustering methodology

The method used for clustering is similar to the one used in an earlier study of collagen research with, however, an important difference.[11] As in the collagen study, the data base was the *Science Citation Index* and the period of cumulation was one year. As before, single-link clustering was used to form clusters of highly cited documents. Two thresholds are needed to unambigously define a cluster: a citation frequency threshold (set for this experiment at 16 citations per document), and a co-citation threshold. The important difference between the present study and the collagen study is the use of a normalized form of the co-citation frequency between the pairs of highly cited documents. If c_i and c_j are the citation frequen-

cies of documents i and j and c_{ij} the frequency of co-citation, a commonly used form of normalization is:

$$\frac{c_{ij}}{c_i + c_j - c_{ij}}$$

which varies from zero to one. This is sometimes referred to as the Jaccard coefficient of similarity.[12] The advantage of this coefficient in clustering highly cited papers is that it eliminates some of the size effect involved in linking documents with widely different citation rates. It also solves the problem of very highly cited method papers, encountered in earlier studies using raw co-citation frequencies as input to clustering.[13] The method papers simply do not cluster at most levels since they are not cited with any other papers a sufficiently high percentage of the time.

A thresholding procedure is used to obtain the single-link clusters.[14] This time, however, a threshold of 0.18 was applied to the normalized co-citation links in the form of the Jaccard coefficient. The links less than 0.18 were then used to determine an association measure between different clusters. One of the remaining questions in this work is the effect of varying thresholds on the structure of the co-citation clusters. For the moment the thresholds of 16 and 0.18 may be considered arbitrary, though based on considerations of machine time practicality and maximum desired cluster sizes.

The same two thresholds are applied to the entire annual *SCI* file which in the case of the 1976 *SCI* resulted in 1928 clusters containing two or more cited documents. These clusters involved a total of 9065 cited documents. A general statistical description of the results of similar annual cluster runs at constant thresholds using the Jaccard coefficient has been presented elsewhere.[15]

After the 1976 cluster run was completed, the recombinant-DNA cluster (#438) was selected for the present study. Its selection was motivated as much by the controversial nature of the subject as by the fact that it was one of the larger clusters in the file (cluster 438 contains 47 highly cited documents and is cited by 475 source papers).

Cluster history

The first question of interest was to trace the history of this cluster in the four previous years of cluster data available to us. The results of this search are presented in Fig. 1 which shows the evolution of what we will call the recombinant-DNA cluster from 1973 (the first year of normalized cluster data available to us) to 1976. The most striking feature is the rapid growth of the cluster from 1975 to 1976 from four documents in 1974, seven in 1975, to 47 in 1976. Two clusters in 1973 (#1243 and #128) each contribute a single document to the

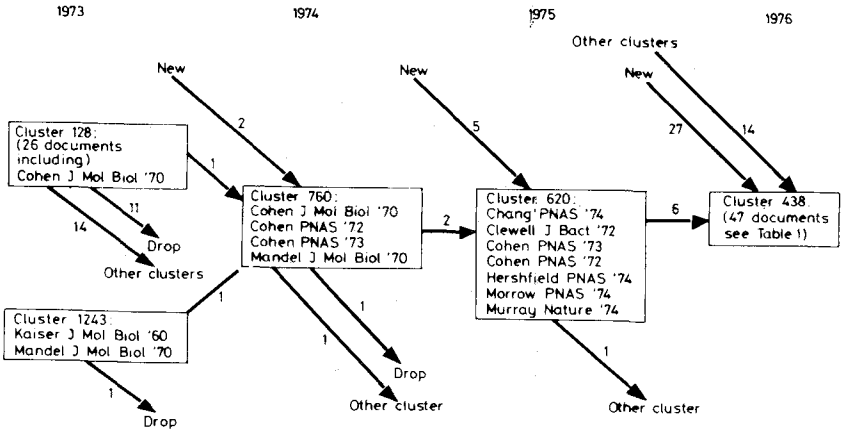

Fig. 1. Recombinant-DNA cluster history

1974 cluster #760, the first real recombinant-DNA cluster. Cluster 1243 consisting of only two documents might be considered the proto-recombinant-DNA cluster, and concerns bacterial transfection. The large 1973 cluster #128 containing 26 documents, in which we find an early paper by Stanley *Cohen* later of recombinant-DNA fame was concerned with drug resistance factor (R-factor) plasmids, and not recombinant-DNA. We will see later that this was *Cohen*'s entree into the recombinant-DNA work. In 1975 there was a large influx of new papers into cluster 620 (five out of seven were new), but none of the papers came from clusters other than #760. Of the 47 documents in 1976, 27 are new in 1976 (they did not appear in any 1975 cluster). Part of the growth of the 1976 cluster is due to its merging with other clusters as represented by the 14 documents originating in other 1975 clusters. We will discuss later what these "other" subjects are, but the important point is that the 1976 cluster represents a convergence of several lines of research, centered around recombinant-DNA. Such a converging pattern is characteristic of rapidly growing specialties in science, as was shown in the previous cluster study of collagen research.

In summary, our cluster history suggests that recombinant-DNA work had its origins in the study of R-factor and transfection in bacteria, and grew rapidly from these origins to the point in 1976 where it became the point of convergence for other related specialties.

Recombinant-DNA cluster

The 47 documents in cluster #438 (each cited 16 times or more) are displayed in Fig. 2 as a connected graph at level 18% of normalized co-citation. The location of the individual documents in the plane was determined by a multi-dimen-

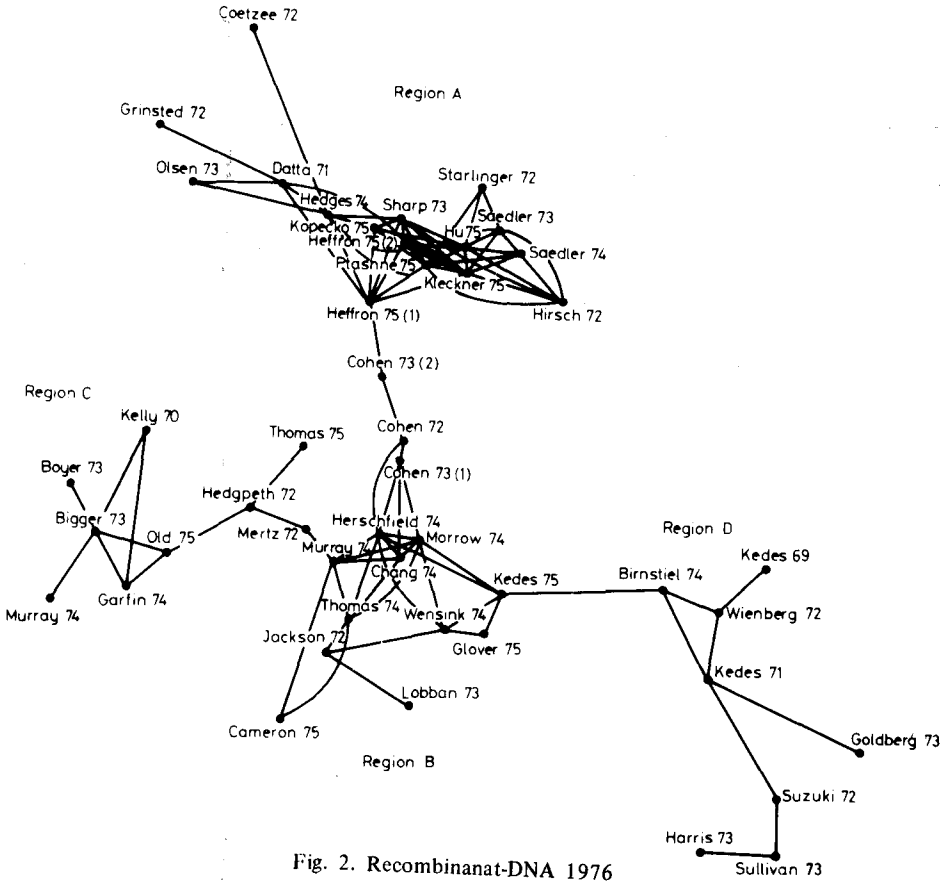

Fig. 2. Recombinanat-DNA 1976

sional scaling analysis of the complete lower-half matrix of normalized co-citation among the 47 documents.[16] High co-citation (here in its normalized form) is interpreted as close proximity, and low co-citation as a large distance of separation. Each node in Fig. 2 has been labeled with the first author of the document and year of publication. These labels are listed in Table 1 along with the full authorship, title and bibliographic description of the document, plus further information to be discussed later.

In order to show the relationship between the scaling plot and the single-link cluster, the inter-document links at the 0.18 threshold have been drawn as undirected lines between points. This illustrates the equivalence of a single-link cluster and a connected graph in graph theory terminology. It also reveals that some parts of the cluster are more densely connected than others and indeed separate sections of the graph are joined together by only single-links at the 0.18 level.

Table 1
Highly cited documents in cluster 438

Document key Context phrase	Percent uniformity	Bibliographic reference
		REGION A
COETZEE 72 plasmid R391(J53)	0.80	J. N. Coetzee, N. Datta, R. W. Hedges, 'R Factors from Proteus rettgeri', *Journal of General Microbiology*, 72 (1972) 543–552.
DATTA 71 plasmid RP4	0.64	N. Datta, R. W. Hedges, E. J. Shaw, R. B. Sykes, M. H. Richmond, 'Properties of an R Factor from Pseudomonas aeruginosa', *Journal of Bacteriology*, 108, No. 3 (Dec. 1971) 1244–2249.
GRINSTED 72 plasmid RP1	1.00	J. Grinsted, J. R. Saunders, L. C. Ingram, R. B. Sykes, M. H. Richmond, 'Properties of an R Factor Which Originated in Pseudomonas aeruginosa 1822', *Journal of Bacteriology*, 110, No. 2 (May 1972) 529–537.
HEDGES 74 translocation of Ap-resistance	0.71	R. W. Hedges, A. E. Jacob, 'Transposition of Ampicillin Resistance deom RP4 to Other Replicons', *Molec. Gen. Genet.*, 132 (1974) 31–40.
HEFFRON 75 (1) translocation of Ap-resistance	0.69	F. Heffron, R. Sublett, R. W. Hedges, A. Jacob, S. Falkow, 'Origin of the TEM Beta-Lactamase Gene Found on Plasmids', *Journal of Bacteriology*, 122, No. 1 (April 1975) 250–256.
HEFFRON 75 (2) TnA inverted repeats	0.86	F. Heffron, C. Rubens, S. Falkow, 'Translocation of a Plasmid DNA Sequence which Mediates Ampicillin Resistance: Molecular Nature and Specificity of Insertion', *Proc. Nat. Acad. Sci. USA*, 72, No. 9 (Sept. 1975) 3623–3627.
HIRSCH 72 polar insertions	0.75	H. Hirsch, P. Starlinger, P. Brachet, 'Two Kinds of Insertions in Bacterial Genes', *Molec. Gen. Genet.*, 119 (1972) 191–206.
HU 75 IS1	0.67	S. Hu, Ohtsubo, N. Davidson, H. Saedler, 'Electron Microscope Heteroduplex Studies of Sequence Relations Among Bacterial Plasmids: Identification and Mapping of the Insertion Sequences IS1 and IS2 in F and R Plasmids', *Journal of Bacterology*, 122, No. 2 (May 1975) 764–775.
KLECKNER 75 translocation of Tc-resistance	0.75	N. Kleckner, R. K. Chan, B.-K. Tye, D. Botstein, 'Mutagenesis by Insertion of a Drug-Resistance Element Carrying an Inverted Repetition', *Journal of Molecular Biology*, 97 (1975) 561–575.
KOPECKO 75 translocation of Ap-resistance	0.55	D. J. Kopecko, S. N. Cohen, 'Site-Specific recA-Independent Recombination Between Bacterial Plasmids: Involvement of Palindromes at the Recombinational Loci', *Proc. Nat. Acad. Sci. USA*, 72 No. 4, (April 1975) 1373–1377.

Table 1 (cont.)

Document key Context phrase	Percent uniformity	Bibliographic reference
OLSEN 73 phage PRRI	0.60	R. H. Olsen, P. Shipley, 'Host Range and Properties of the Pseudomonas aeruginosa R Factor R1822', *Journal of Bacteriology*, 113, No. 2 (Feb. 1973) 772–780.
PTASHNE 75 R-plasmid insertion sequences	0.92	K. Ptashne, S. N. Cohen, 'Occurrence of Insertion Sequence (IS) Regions on Plasmid Deoxyribonucleic Acid as Direct and Inverted Nucleotide Sequence Duplications', *Journal of Bacteriology*, 122, No. 2 (May 1975) 776–781.
SAEDLER 74 IS2	1.00	H. Saedler, H. J. Reif, S. Hu, N. Davidson, 'IS2, A genetic Element for Turn-off and Turn-on of Genetic Activity in E. coli', *Molec. Gen. Genet.*, 132 (1974) 265–289.
SAEDLER 73 insertion sequences	1.00	H. Saedler, B. Heiss, 'Multiple Copies of the Insertion-DNA Sequences IS1 and IS2 in the Chromosome of E. coli K-12', *Molec. Gen. Genet.*, 122 (1973) 267–277.
SHARP 73 drug resistance at inverted repeats	0.67	P. A. Sharp, S. N. Cohen, N. Davidson, 'Electron Microscope Heteroduplex Studies of Sequence Relations among Plasmids of Escherichia coli – II. Structure of Drug Resistance (R) Factors and F. Factors', *Journal of Molecular Biology*, 75 (1973) 235–255.
STARLINGER 72 insertion sequences	0.83	P. Starlinger, H. Saedler, 'Insertion Mutations in Microorganisms', *Biochimie*, 54 (1972) 177–185.
		REGION B
CAMERON 75 bacteriophage λ vector	1.00	R. Cameron, S. M. Panasenko, I. R. Lehman, R. Davis, 'In Vitro Construction of Bacteriophage λ Carrying segments of the Escherichia coli chromosome: Selection of Hybrids Containing the Gene for DNA Ligase', *Proc. Nat. Acad. Sci. USA*, 72, No. 9 (Sept. 1975) 3416–3420.
CHANG 74 staphylococcus plasmid	0.30	A. C. Y. Chang, S. N. Cohen, 'Genome Construction Between Bacterial Species In Vitro: Replication and Expression of Staphylococcus Plasmid Genes in Escherichia coli', *Proc. Nat. Acad. Sci. USA*, 71, No. 4 (April 1974) 1030–1034.
COHEN 72 transformation	0.87	S. N. Cohen, A. C. Y. Chang, L. Hsu, 'Nonchromosomal Antibiotic Resistance in Bacteria: Genetic Transformation of Escherchia coli by R-Factor DNA', *Proc. Nat. Acad. Sci. USA*, 69, No. 8 (Aug. 1972) 2110–2114.

Table 1 (cont.)

Document key Context phrase	Percent uniformity	Bibliographic reference
COHEN 73 (1) plasmids	0.82	S. N. Cohen, A. C. Y. Chang, H. W. Boyer, R. B. Helling, 'Construction of Biologically Functional Bacterial Plasmids In Vitro', *Proc. Nat. Acad. Sci. USA*, 70, No. 11 (Nov. 1973) 3240–3244.
COHEN 73 (2) pSC101	0.70	S. N. Cohen, A. C. Y. Chang, 'Recircularization and Autonomus Replication of a Sheared R-Factor DNA Segment in Escherichia coli Transformants', *Proc. Nat. Acad. Sci. USA*, 70, No. 5 (May 1973) 1293–1297.
GLOVER 75 drosophila DNA cloning	0.63	D. M. Glover, R. L. White, D. J. Finnegan, D. S. Hogness, 'Characterization of Six Cloned DNAs from Drosophila melanogaster, Including one that contains the Genes for rRNA', *Cell*, 5 (June 1975) 149–157.
HERSHFIELD 74 Co1E1	0.56	V. Hershfield, H. W. Boyer, C. Yanofsky, M. A. Lovett, D. R. Helinski, 'Plasmid Co1E1 as a Molecular Vehicle for Cloning and Amplification of DNA', *Proc. Nat. Acad. Sci. USA*, 71, No. 9 (Sept. 1974) 3455–3459.
JACKSON 72 DNA joining	0.63	D. A. Jackson, R. H. Symons, P. Berg, 'Biochemical Method for Inserting New Genetic Infromation into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli', *Proc. Nat. Acad. Sci. USA*, 69, No. 10 (Oct. 1972) 2904–2909.
KEDES 75 histone DNA cloning	0.56	L. H. Kedes, A. C. Y. Chang, D. Houseman, S. N. Cohen, 'Isolation of Histone Genes from Unfractionated Sea Urchin DNA by Subculture Cloning in E. coli', *Nature*, 255 (June 1975) 533.
LOBBAN 73 terminal transferase	0.75	P. E. Lobban, A. D. Kaiser, 'Enzymatic End-to-End Joining of DNA Molecules', *J. Mol. Biol.*, 78 (1973) 453–471.
MORROW 74 frog DNA cloning	0.33	J. F. Morrow, S. N. Cohen, A. C. Y. Chang, H. W. Boyer, H. M. Goodman, R. B. Helling, 'Replication and Transcripton of Eukaryotic DNA in Escherichia coli', *Proc. Nat. Acad. Sci. USA*, 71, No. 5 (May 1974) 1743–1747.
MURRAY 74 bacteriophage λ vector	0.67	N. E. Murray, K. Murray, 'Manipulation of Restriction Targets in Phage λ to Form Receptor Chromosomes for DNA Fragments', *Nature*, 251 (Oct. 11, 1974) 476.
THOMAS 74 bacteriophage λ vector	0.82	M. Thomas, J. R. Cameron, R. W. Davis, 'Viable Molecular Hybrids of Bacteriophage Lambda and Eukaryotic DNA', *Proc. Nat. Acad. Sci. USA*, 71, No. 11 (Nov. 1974) 4579–4583.

Table 1 (cont.)

Document key Context phrase	Percent uniformity	Bibliographic reference
WENSINK 74 drosophila DNA cloning	0.50	P. C. Wensink, D. J. Finnegan, J. E. Donelson, D. S. Hogness, 'A System for Mapping DNA Sequences in the Chromosomes of Drosophila melanogaster', *Cell*, 3 (Dec. 1974) 315–325.
		REGION C
BIGGER 73 restriction endonuclease recognition sequences	0.60	C. H. Bigger, K. Murray, N. E. Murray, 'Recognition Sequence of a Restriction Enzyme', *Nature New Biology*, 244 (July 4, 1973) 7.
BOYER 73 restriction endonuclease recognition sequences	0.86	H. W. Boyer, L. T. Chow, A. Dugaiczyk, J. Hedgpeth, H. M. Goodman, 'DNA Substrate Site for the Eco_{RII} Restriction Endonuclease and Modification Methylase', *Nature New Biology*, 244 (July 1973) 40.
GARFIN 74 Hpa I and II	0.57	D. E. Garfin, H. M. Goodman, 'Nucleotide Sequences at the Cleavage Sites of Two Restriction Endonucleases from Hemophilus Parainfluenzae', *Biochemical and Biophysical Research Communication*, 59, No. 1 (1974) 108.
HEDGPETH 72 DNA cleavage by EcoR1	0.77	J. Hedgpeth, H. M. Goodman, H. W. Boyer, 'DNA Nucleotide Sequence Restricted by the RI Endonuclease', *Proc. Nat. Acad. Sci. USA*, 69, No. 11, (Nov. 1972) 3448–3452.
KELLY 70 restriction endonuclease recognition sequences	1.00	T. J. Kelly, Jr., H. O. Smith, 'A Restriction Enzyme from Hemophilus influencae II. Base Sequence of the Recognition Site', *J. Mol. Biol.*, 51 (1970) 393–409.
MERTZ 72 DNA cleavage by EcoRI	0.80	J. E. Mertz, R. W. Davis, 'Cleavage of DNA by R_1 Restriction Endonuclease Generates Cohesive Ends', *Proc. Nat. Acad. Sci. USA*, 69, No. 11 (Nov. 1972) 3370–3374.
MURRAY 74 restriction endonuclease	1.00	K. Murray, R. W. Old, 'The Primary Structure of DNA', *Progr. Nucleic. Acid. R.*, 14 (1974) 117.
OLD 75 Hind III	0.75	R. Old, K. Murray, 'Recognition Sequence of Restriction Endonuclease III from Hemophilus influenzae', *J. Mol. Biol.*, 92 (1975) 331–339.
THOMAS 75 EcoRI fragments of bacteriophage λ	0.86	M. Thomas, R. W. Davis, 'Studies on the Cleavage of Bacteriophage Lambda DNA with EcoRI Restriction Endonuclease', *J. Mol. Biol.*, 91 (1975) 315–328.

Table 1 (cont.)

Document key Context phrase	Percent uniformity	Bibliographic reference
		REGION D
BIRNSTIEL 74 histone gene reiteration	0.86	M. L. Birnstiel, J. Telford, E. Weinberg, D. Stafford, 'Isolation and Some Properties of the Genes Coding for Histone Proteins', *Proc. Nat. Acad. Sci. USA*, 71, No. 7 (July 1974) 2900–2904.
GOLDBERG 73 mRNA sequence repetition	0.75	R. B. Goldberg, G. A. Galau, R. J. Britten, E. H. Davidson, 'Nonrepetitive DNA Sequence Representation in Sea Urchin Embryo Messenger RNA', *Proc. Nat. Acad. Sci. USA*, 70, No. 12, (Dec. 1973) 3516–3520.
HARRIS 73 non-repetitive ovalbumin mRNA	0.60	S. E. Harris, A. R. Means, W. M. Mitchell, B. W. O'Malley, 'Synthesis of (^{3}H)DNA Complementary to Ovalbumin Messenger RNA: Evidence for Limited Copies of the Ovalbumin Gene in Chick Oviduct', *Proc. Nat. Acad. Sci. USA*, 70, No. 12, Part II (Dec. 1973) 3776–3780.
KEDES 69 sea urchin histone mRNA	1.00	L. H. Kedes, P. R. Gross, 'Identification in Cleaving Embryos of Three RNA Species Serving as Templates for Synthesis of Nuclear Proteins', *Nature*, 223 (1969) 1335.
KEDES 71 histone gene reiteration	1.00	L. H. Kedes, M. L. Birnstiel, 'Reiteration and Clustering of DNA Sequences Complementary to Histone Messenger RNA', *Nature New Biology*, 230 (April 7, 1971) 165.
SULLIVAN 73 non-repetitive ovalbumin mRNA	0.80	D. Sullivan, R. Palacios, J. Stavnezer, J. M. Taylor, A. J. Faras, M. L. Kiely, N. M. Summers, J. M. Bishop, R. T. Schimke, 'Synthesis of a Deoxyribonucleic Acid Sequence Complementary to Ovalbumin Messenger Ribonucleic Acid and Quantification of Ovalbumin Genes', *The Journal of Biological Chemistry*, 248, No. 21 (Nov. 10, 1973) 7530–7539.
SUZUKI 72 non-repetitive fibroin mRNA	0.83	Y. Suzuki, L. P. Gage, D. D. Brown, 'The Genes for Silk Fibroin on Bombyx mori', *J. Mol. Biol.*, 70 (1972) 637–649.
WEINBERG 72 sea urchin histone gene reiteration	1.00	E. S. Weinberg, M. L. Birnstiel, I. F. Purdom, R. Williamson, 'Genes Coding for Polysomal 9S RNA of Sea Urchins: Conservation and Divergence', *Nature*, 240 (Nov. 24, 1972) 225.

A visual inspection of Fig. 2 suggests division of the network into four regions, which we have labeled A, B, C and D. An examination of the titles of the papers in Table 1 suggests the following regional designations:

Region A: translocation and insertion of drug resistant genes,

Region B: recombinant-DNA,

Region C: restriction endonucleases,

Region D: repetitive and non-repetitive gene sequences.

Regions A, C, and D are the previously separate clusters which converged around recombinant-DNA, region B, in 1976. Note that the three regions, are each connected to B by a single linkage. It is quite clear in this case that the structure of the co-citation cluster is most easily explained in terms of distinct, but related, subject matters.

At first glance our representation of the specialty with documents of varying ages may seem ahistorical. The cluster is derived for a single "moment" in time (a single year's cumulative picture), in the manner of a snapshot. The "events" represented by documents in the cluster should be seen as coexisting, in some collective sense, for a group of scientists, perhaps representing a prevailing paradigm or set of exemplars for the research area, much in the sense that *Kuhn* intended it.[17] That such a cross-sectional view is a mixture of events and findings of varying ages is analogous to the snapshop of a room in which the items of furniture have been placed at various times in the past, but all co-exist at the moment the picture is snapped. Our interpretation of the co-citation cluster follows as if we were exploring the "mental furniture" of a hypothetical and ideal recombinant-DNA researcher.

Citation context analysis

After obtaining a spatial display of the documents in the cluster, the next step is to determine how the subject matter varies over the cluster map. Clearly the highly cited papers' titles can assist us here, but this cannot tell us what the citing authors thought of the papers when they cited them singly or jointly. Therefore it seems that we must return to the citing papers themselves to understand the significance of the cluster. In effect, we are not concerned with the clustered documents as we see them, but rather as the citing authors have seem them. What is important is what the cited documents symbolize for the citing authors.

To explore this symbolic reality, we must examine the text of the citing paper and analyze the point or points where reference was made to the document in question. After examining several such contexts for different citing authors, a pattern of common usage may begin to emerge and we can estimate the degree of

commonality of usage by calculating the number of equivalent usages divided by the total sampled. We will call this statistic, as it was called in an earlier paper, the percent uniformity. A sample of about nine citing papers was obtained for each of the cited documents on the cluster map (a total of 414 contexts was examined for the 47 documents for an average of 8.8 contexts per document). The sentences around the footnote number corresponding to the cited document are recorded, and then compared with one another. The concept or idea expressed by the largest number of citation contexts is determined and each context is coded for its presence or absence. The ratio of the number of contexts expressing the most prevalent concept to the total number examined (the percent uniformity) is then calculated.

The results of the citation context analysis are presented in two ways: First in Table 1 we have listed each of the documents by region, giving the word or phrase most frequently associated with each document in the citation contexts, and the percent uniformity of that usage. Second, in Figs 3a, b, c, and d, a blow-up of each of the regions is provided, substituting the context phrase for the author-year designation, and giving the percent uniformity.

Turning first to region A (Fig. 3a), which we have called "translocation and insertion of drug resistant genes", we can distinguish roughly three somewhat overlapping subregions. On the right, the triangularly shaped group of document-concepts concerns insertion sequences (IS = insertion sequence). An insertion sequence is a piece of DNA which can be incorporated at various locations on the bacterial chromosome or move between the chromosome and bacterial plasmids and viruses. Insertion sequences can cause mutations and can undergo natural recombination with each other. The middle and rather linear sub-region of Region A concerns the translocation of various kinds of antibiotic resistant genes (R-factors). Translocation is a naturally occurring process whereby segments of DNA (in this case from R-factors) move from one plasmid to another, and from one bacterium to another, the mobile piece being the insertion sequence. The practical concern here is the spread of drug resistance in various bacteria due to the transferring of R-factor genes among plasmids that inhabit bacteria. Finally a rather diffusely connected subregion on the left of region A consisting of four documents is concerned with antibiotic resistance genes or R-factors (plasmids and bacteriophage) which are used in translocation experiments. We will see that plasmids and phages also play a prominent role in recombinant-DNA (Region B), although different DNA vectors are involved. In a sense, the R-factors were the parents of the cloning vehicles used in the recombinant-DNA work, and hence provided input to that development, although the area represented by region A is an independent development.

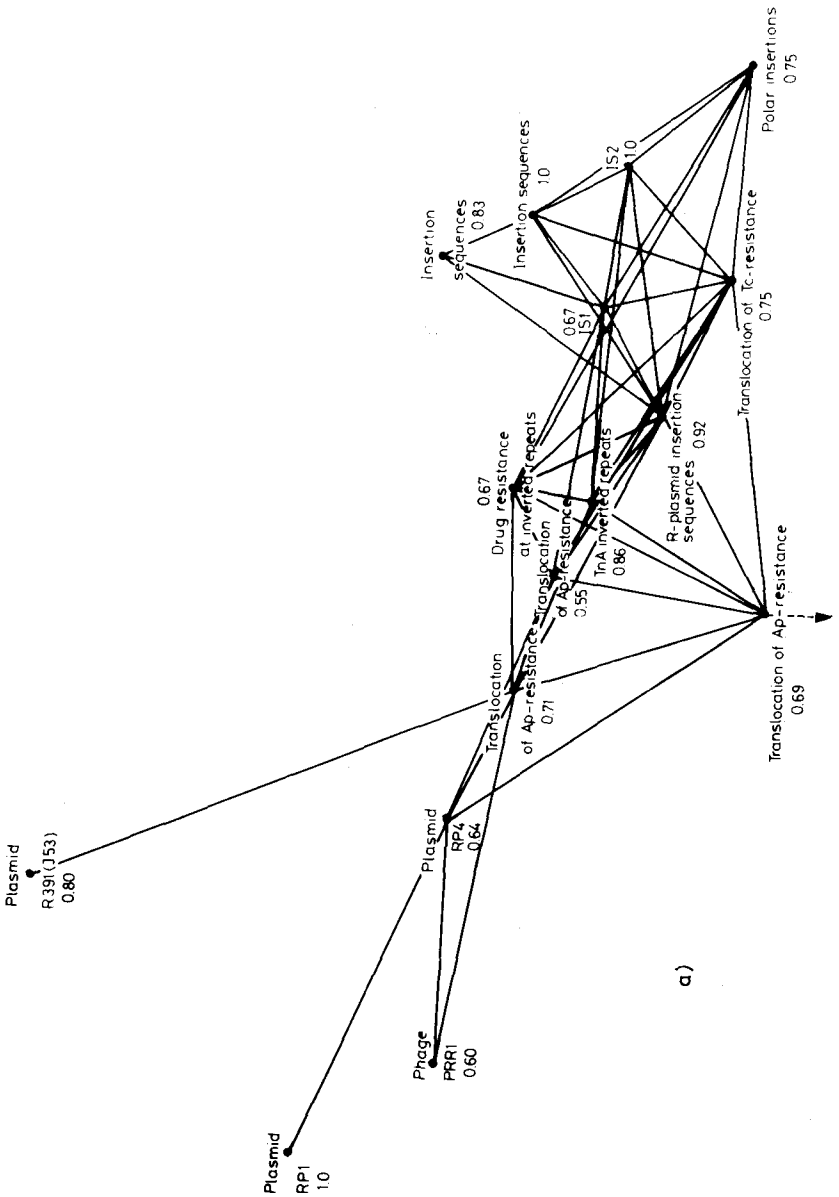

Fig. 3a. Region A

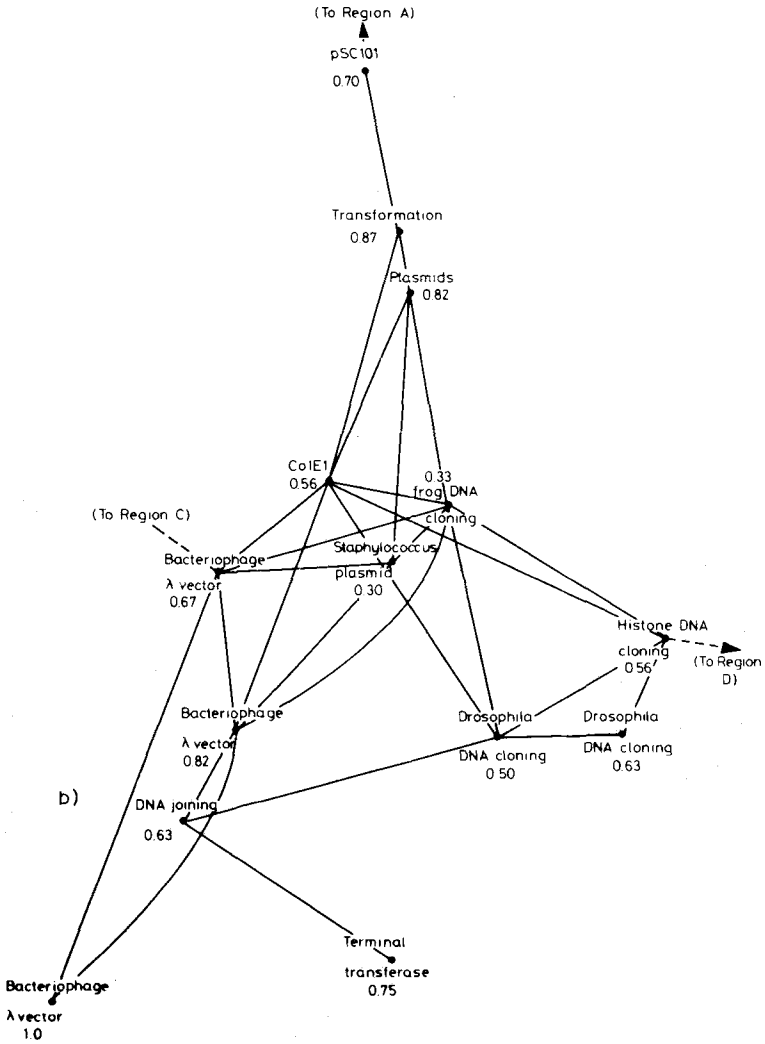

Fig. 3b. Region B

Region A is tenuously linked to region B, recombinant-DNA proper, via *Cohen 73(2)* (see Fig. 3b). This document is associated with the well-known pSC101 (plasmid Stanley *Cohen 101*) which was the first plasmid to be successfully employed in a recombinant-DNA experiment involving the introduction and propagation of foreign DNA in *E-coli*. We will return to the nature of this bridge between regions A and B later on, when we consider the data on co-citation context analysis.

Region B has a somewhat more complex sub-structure than region A. Roughly three categories of document-concepts appear: (1) those associated with different types of vectors (plasmids or phages) used to carry the foreign DNA into the host *(E-coli);* (2) those representing various species whose DNA has been successfully propagated in *E-coli*; and (3) various methodologies for joining DNA fragments and for introducing foreign DNA into *E-coli*. This last named methodology called "transformation", involves treatment with calcium chloride and heat shock, and is associated with *Cohen 72*. It is located near the top of region B, and is clearly a key method in the specialty. Two other methods are located near the bottom of region B: DNA joining and terminal transferase. Both are used for joining DNA fragments with the enzyme terminal transferase, and predate the recombinant-DNA techniques which were later to exploit special features of the restriction endonucleases.

Five papers in the region are associated with the cloning of DNA from various species in *E-coli* using recombinant-DNA techniques. These form a fairly coherent sub-group to the right of center. The species involved are: a bacterium *(Staphylococcus aureus)*; toad *(Xenopus laevis)*; fruit fly *(Drosophila)*; and sea urchin (labeled "histone DNA cloning"). The toad cloning work was the first successful propagation of eukaryotic genes in a prokaryotic organism.

The six remaining papers in region B stand for a variety of cloning vectors, including the previously mentioned pSC101 plasmid. Three stand for the mutant *E-coli* virus called bacteriophage lambda; one for plasmids in general (of which more later); one for a plasmid called ColE1 (Colicinogenic factor E1); and the last for pSC101. These papers are not grouped in a compact sub-region but are spread vertically in a roughly linear fashion, intermixing with the other papers of region B. The three bacteriophage papers do, however, group at the lower end of the region.

Region C (Fig. 3c) is connected to region B via *Mertz 72* which is associated with the study of DNA cleavage by a restriction enzyme called EcoR1. The special properties of this enzyme and the specific way it cut DNA were exploited in the recombinant-DNA work represented in region B. Hence the linkage between the regions is one of historical as well as operational dependence.

Region C as a whole concerns restriction endonucleases. Restriction endonucleases are enzymes which cleave the DNA molecule at highly specific points. At the left-most part of the region are four papers associated with the ability of these enzymes to recognize certain nucleotide sequences on DNA and cleave at these highly specific points. It turned out that the enzymes cleaved DNA at or near an axis of rotational symmetry, resulting in the well-known palindrome structure of the cleavage sites. Toward the center of region C are two papers associated with the specific restriction enzymes, Hpa and Hind III. These enzymes are drived from the *Haemophilous* bacte-

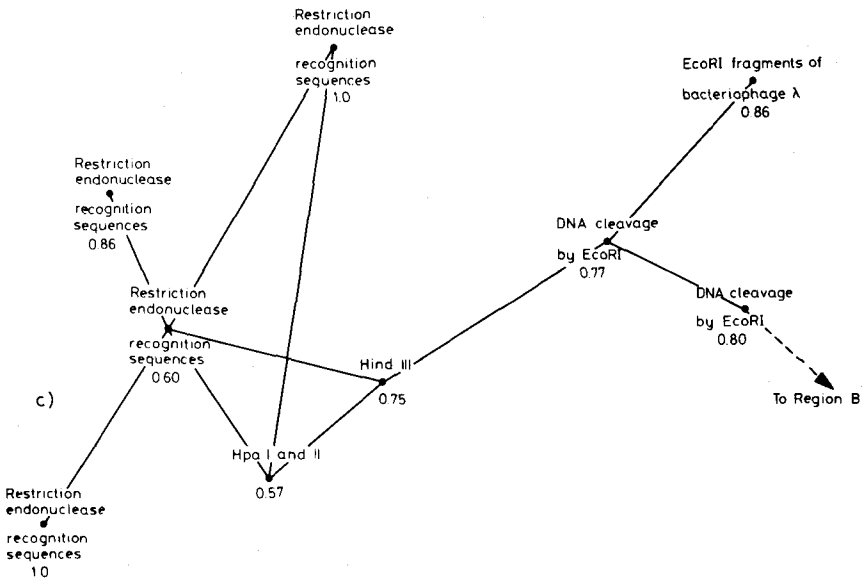

Fig. 3c. Region C

rium, rather than *E-coli,* from which EcoR1 is derived. Three papers associated with EcoR1 appear at the right of region C, consistent with the important role this enzyme plays in recombinant-DNA experiments. EcoR1 has the special ability to cleave DNA so that the ends are "cohesive", that is, capable of being rejoined with other DNA fragments with similar "cohesive" ends. This made the biochemical procedure of rejoining DNA molecules much less complicated.

Turning to Region D (Fig. 3d), we have a more purely genetic concern and one which is more a user of recombinant-DNA technology than a source of that technology. The point of attachment with region B is *Birnstiel 74* which is associated with the study of gene reiteration on sea urchin histone DNA. The region can be divided into two sub-regions, one having to do with the study of repeated or reiterated gene sequences (mainly in histone m-RNA) and the other with non-repetitive sequences (in fibroin and ovalbumin). The basic concern of this research, beyond understanding why certain genes occur in single copies and others in multiple copies, is to understand the regulation of gene expression. Papers in this region do not utilize recombinant-DNA techniques, but more classical procedures. The significance of the attachment of this region to region B is that with the advent of recombinant-DNA techniques, the experimental problem of studying gene reiteration was greatly simplified. Hence the link is one between an established problem area in genetics and a new set of methodological tools applicable to the

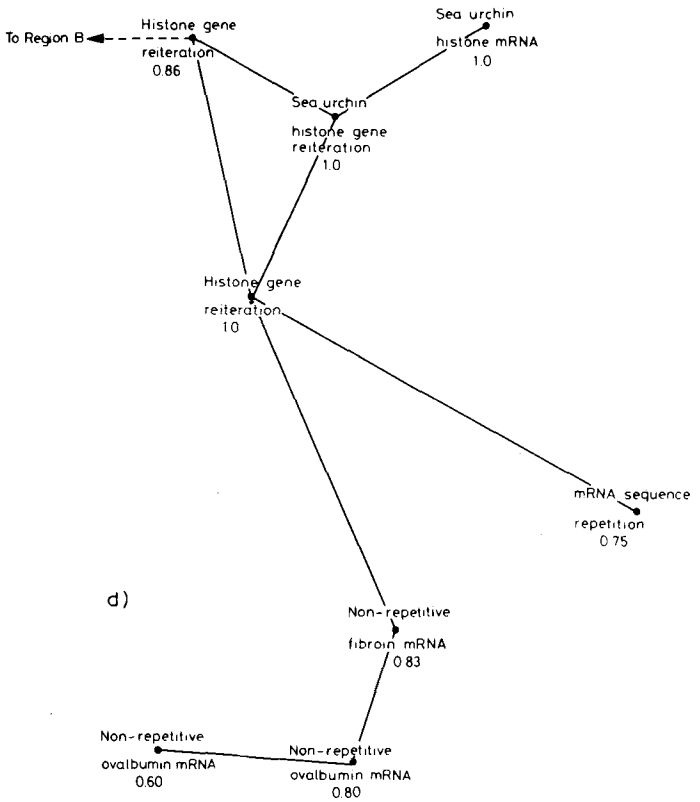

Fig. 3d. Region D

problem. In fact, *Kedes 75* which links region D to region B specifically applies recombinant-DNA methods to the same problem of histone gene reiteration as *Birnsteil 74* does using more classical procedures.

General characteristics

Before delving more deeply into the cognitive structure implied by the map, we turn to some general characteristics. First, different densities of linkage are apparent in different regions, with region A having the highest density, region B next highest, and C and D least. The reason for these differences is by no means clear although in this case there is a slight tendency for the more dense regions to contain papers of a more recent origin. Highly dense patterns of co-citation indicate a tendency of authors to cite many of the papers in the region together, and may be more characteristic of a youthful specialty.

Of the 47 document-concepts on the map, roughly 28 or 60% of them refer to objects or entities usually of a molecular sort, while 19 or 40% of them refer to processes of phenomena. There appear to be no significant variation in these proportions across regions. This tells us simply that recombinant-DNA research is very much concerned with the behaviour and manipulation of molecular units. Another more interesting observation is that some of the nodes have been labeled with the same concept. This means that citing authors associated the same concepts with different documents. This phenomenon of "redundancy" of citation was noted by *Moravcsik*. There are a total of eight groups of redundant nodes on the maps: One group of four redundant documents; two groups of three documents; and five groups of two documents. The redundant groups are distributed over all regions: each of the four regions (A, B, C, and D) has two redundant groups. Most significant, however, is that redundant documents are almost always one co-citation step removed from one another, i.e. redundant documents are frequently co-cited. This feature was observed in an earlier study of highly cited chemistry documents.

Of course, this redundancy results in most cases from authors' citing two or more works at the same point in their texts suggesting to the reader that the cited works are equivalent in some sense. Our redundant documents are not quite structurally equivalent since they still may link to different non-redundant documents. Redundancy may arise from independent co-discovery, replication or corroboration by different scientists, or repetitive publications by the same scientist.

Other general features of the map involve the statistic we have called percent uniformity, the relative frequency that the most prevalent concept is associated with the cited work. The overall mean uniformity for documents in the cluster is 76% which is considerably lower than the 87% mean uniformity found previously for a sample of very highly cited chemistry documents. It was hypothesized in that earlier study that the uniformity for research front papers like the recombinant-DNA papers would be lower due to the short time available for consensus to be reached on the documents' meaning in the specialty community. The data from cluster 438 confirm this lower uniformity for research front papers.

To examine the lag time in consensus formation more closely we selected a single document from the recombination-DNA cluster, *Cohen* 1972 on the transformation method, and looked up all citations to it back to 1972 in the *Science Citation Index*. The paper has received a total of 57 citations up to and including 1976. An examination of the contexts of each of these citations reveals remarkably high uniformity (0.80) beginning with the earliest references during 1972 and 1973. The uniformity increased to 0.87 by 1976 (the 1972 – 1974 uniformity was 0.77, while the 1975 – 1976 uniformity was 0.84). The increase is not dramatic, and what is more striking is the early date at which the uniform citing of

this document as the "transformation" method emerged. This suggests that the significance of this paper was established in the prepublication stage, perhaps through pre-print distribution or discussion at meetings. Whatever the reasons, we have not observed significant time delays associated with a negotiation of meaning by the specialty community.

Examination of the distribution of the uniformity values over the map reveals another pattern of interest. There is a tendency for documents at the periphery of the cluster and those linked rather tenuously by one or two links to have higher uniformity than documents at the centers of the regions having many links. This tendency is also reflected in higher mean uniformity for the sparsely connected regions C and D and lower uniformities for the more highly interconnected regions B and A. A possible explanation for this is that documents at the center of regions are more variable in their meaning (multi-valent) depending upon which other documents they are co-cited with, and thus a lower percentage of authors associate them with a particular concept.

The kind of analysis required to study the possible multi-valent character of highly cited papers is to separate citations contexts associated with the different co-citation links. For example, is the pSC101 paper, *Cohen 73(2),* cited differently when it is co-cited with *Heffron 75(1)* than when it is cited with *Cohen 72?* This question remains for future research.

Co-citation context analysis

The previous discussion leads us to consider the next step in citation context analysis, that is, its extension to co-citation contexts. The question is whether the co-citation link itself corresponds to some standardized usage, for example, whether it specifies that two concepts represented by the co-cited documents bear a certain relationship to one another. Rather than examining the text at the point a footnote number appears, co-citation context analysis locates text *in between* the citations to two different documents. If these are separated by a large amount of text, the analysis becomes difficult and requires "reading between the lines". If, however, the two documents are cited in close proximity (though they are not redundant), an analysis of the co-citation context becomes feasible.

We have used this procedure to examine the links on the recombinant-DNA map which form the bridges from one region to another. The purpose of the analysis was to understand why the regions were linked together. First consider the bridge between regions B and A, formed by the *Cohen 73(2) – Heffron 75(1)* link. We recall that *Cohen 73(2)* is associated with the plasmid pSC101 while *Heffron 75(1)* is associated with translocation of Ap-resistance. Analysis of five co-citation contexts reveals that the connection here is due to the fact that

pSC101 has the ability to undergo insertion of Ap-resistant genes. The plasmid has recipient sites for these genes, and these genes can, furthermore, be translocated from one pSC101 plasmid to another. Hence, the bond between regions B and A is between a physical entity (a plasmid) and a phenomenon (translocation of Ap-resistance) to which it is prone.

The second case is the bridge between regions B and C, involving the link between *Murray 74* (bacteriophage lambda as vector) and *Mertz 72* (DNA cleavage by EcoRI). Eight co-citation contexts were examined and the relationship implied here is one of temporal order: *first* DNA is cleaved by EcoRI and *then* it is inserted into bacteriophage lambda. This case might be regarded as a procedural ordering, successive steps in a recombinant-DNA experiment.

The third case involves the link between *Kedes 75* (histone DNA cloning) in region B and *Birnsteil 74* (histone gene reiteration) in region D. The relationship here is that the recombinant-DNA or cloning methods used in *Kedes 75* are a new way of studying histone gene reiteration, which was approached in a more classical way in *Birnsteil 74*. Hence, the bond consists in a common concern with understanding the phenomenon of gene reiteration. It is a link between a phenomenon and a technique which can be used to study it.

We have not examined the contexts of co-citation beyond these three bridging cases. Yet it is clear that the approach could be used to study each of the linkages on the map and could shed light on the reasons the documents are connected, i.e., the logic which holds the cluster of concepts together.

Validation

Assuming that cluster 438 represents a snapshot or stop-action of recombinant-DNA research, to test our model we need a contemporary account which sets out a more or less official statement of the history and current state of the field circa 1976. Such a statement is the *Scientific American* article by Stanley Cohen in 1975 entitled "The Manipulation of Genes".[18]

We will use *Cohen*'s article as a guide through region B of our map. The first development discussed by *Cohen* is the discovery of DNA ligases in 1967, enzymes capable of repairing breaks in DNA strands. None of the documents in cluster 438 deal specifically with DNA ligases and in fact, the oldest paper on the map is from 1969. Hence this part of recombinant-DNA prehistory is not visible in our 1976 data. The second development described by *Cohen*, however, is. He notes that in order for DNA ligase to act efficiently, a method for holding the two ends of the DNA together is required. The work of *Lobban* and *Kaiser (Lobban 73)* and *Jackson, Symons* and *Berg (Jackson 72)*, both groups from Stanford,

accomplished this independently using the enzyme terminal-transferase. Papers associated with this enzymatic joining procedure are located at the bottom of region B. *Jackson 72* (with Paul *Berg* as co-author) is often regarded as a seminal work which showed that it was possible to construct a cloning vehicle.

At this point in the story *Cohen* introduces the restriction endonucleases, because of their ability to make cohesive termini for joining DNA strands, a much simpler joining procedure than using terminal-transferase. He mentions the recognition sequence work (at the left of region C) using restriction endonucleases in the early 1970's, and the work on the EcoR1 restriction enzyme at the right of region C which created cohesive termini. Two of the three findings which were mentioned by *Cohen* relating to the special properties of EcoR1 are represented in the cluster. The *Mertz* and *Davis* paper of 1972 and the *Hedgpeth, Goodman,* and *Boyer* paper of 1972 are in the cluster; but the work of *Sgaramella* is not.[19] The last named paper was cited 12 times in the 1976 *SCI* and thus fell short of our cut-off of 16 citations used for clustering.

Next *Cohen* takes up the problem of how foreign DNA is introduced into an organism such as *E-coli* where it may be propagated. He discusses the development of the transformation technique which corresponds to *Cohen 72* near the top of region B. As a prelude to this he mentions prior work with plasmids which carried antibiotic resistance (R-factors), a field represented by region A on the map. As we saw from cluster history, for *Cohen* at least, the route to recombinant-DNA was through the study of R-factor plasmids in bacteria. A precursor to *Cohen*'s transformation method was work by *Mandel* and *Higa* which showed that *E-coli* treated with calcium salts could take up viral DNA.[20] The *Mandel* paper appears in the 1973 and 1974 recombinant-DNA clusters shown in Fig. 1, but not in the 1976 cluster. Following this development, *Cohen, Chang* and *Hsu* in 1972 showed that *E-coli* could be made permeable to R-factor plasmids and replicate there. This is the *Cohen 72* transformation paper on the map.

Cohen then discusses the development of plasmids as vehicles for introducing and propagating foreign DNA in *E-coli*. He mentions his continuing work with R-factor plasmid and his attempts to mechanically shear the R-factor plasmid and then introduce fragments of it into *E-coli* by tranformation. He and *Chang* found in the process a plasmid, later called pSC101, which would become important in later experiments. The paper reporting the creation of this plasmid by shearing and transformation of *E-coli* is the one associated with pSC101 [*Cohen 73(2)*] on our map, the bridging paper for region A and B.

Shortly after these shearing experiments, they began using restriction endonucleases and in particular EcoR1, exploiting its newly discovered property of creating "sticky ends", for cleaving plasmid DNA. At this time a collaboration

between *Cohen* and *Chang* of Stanford and *Boyer* and *Helling* of University of California – San Francisco began which resulted in the *Cohen 73(1)* paper. They found that one of the plasmids in *Cohen*'s possession, which they referred to as pSC101, has the special virtue of having only one cleavage site for EcoR1, and that the plasmid after introduction into *E-coli* by transformation retained its tetracycline resistance. In *Cohen 73(1)* they reported the first successful introduction of a foreign DNA fragment, from another *E-coli* plasmid carrying kanamycin resistance, into *E-coli* using pSC101 as the carrier. Hence, plasmids had a two-fold significance for this paper: a plasmid was used as a vector to carry foreign DNA and another plasmid was used as the source of foreign DNA.

The next step described by *Cohen* was his work with *Chang* to show that genes from another species could be introduced into a plasmid and could transform *E-coli*. [The two plasmids used in *Cohen 73(1)* were both native to *E-coli*]. The resulting paper, *Chang 74* on our map, showed how genes for penicillin resistance from a plasmid native to the bacterium *Staphylococcus aureus* could be introduced into *E-coli* and function there. This was the first time the species barrier had been broken by the transfer of genetic information between unrelated organisms. In logical progression, *Cohen* then describes the breaching of the eukaryotic/prokaryotic barrier. This was the work of *Morrow, Cohen, Chang, Boyer, Goodman* and *Helling* (*Morrow 74*, on our map). The eukaryotic genes used were from species *Xenopus laevis*, the South African clawed toad. They found that the animal genes could be introduced into *E-coli* using the same procedures as before: EcoR1 cleavage, a plasmid vector, and transformation.

Cohen briefly mentions other developments that occurred in quick succession after these 1974 experiments. He notes that other plasmids were soon found having desirable properties like the original pSC101. One of these is most certainly the ColE1 plasmid associated with *Herschfield 74* on our map, although *Cohen* does not mention it specifically.[21] ColE1 is noteworthy for its ability to amplify DNA spliced to it, greatly facilitating purification of the cloned DNA. He mentions the development of bacteriophage lambda as a cloning vehicle by researchers in Edinburgh. We have, of course, three redundant papers in region B associated with bacteriophage lambda. To illustrate how the new cloning techniques can be used to investigate the complex genetic make-up of higher organisms he mentions the fruit fly *(Drosophila)* work corresponding to *Wensink 74* and *Glover 75* at the lower right of region B. The only work in region B he does not refer to directly is the sea urchin histone DNA cloning of *Kedes 75*[22] which was published the same year that he wrote his *Scientific American* article.

We see that the *Cohen* article provides a nearly complete guide to region B and adjacent parts of other regions, and only a few developments he mentions are not

represented on the map. One general conclusion is that the structure of region B bears little relationship to the chronological sequence of events in the field. The structure seems more determined by ahistorical subject similarity together with logical or procedural relationships. This supports our interpretation of the map, not as a "history" of recombinant-DNA but as a snapshot of the paradigm for the field circa 1976.

Conclusions

Our goal has been to see what can be learned about the field of recombinant-DNA using the co-citation clustering method. We have combined this methodology with citation context analysis to provide a detailed cognitive structure for the research area. This paper is a first attempt to combine these two methodologies.

The combination of co-citation clustering and citation context analysis has revealed that some of the highly cited documents on the map are redundant, that is, associated with the same concepts. Redundant nodes were found to be very close to one another on the map. We found a higher uniformity of usage for documents on the periphery of the cluster, having few links to other documents, than for centrally located and highly connected documents. This suggests that the number of links a document has to other documents is related to the number of use and meanings the document has.

Another finding was the rapidity with which a paper achieves its most prevalent mode of use or meaning. *Cohen*'s *72* paper achieved a high level of uniformity of usage (80%) in the first year after publication, and the uniformity increased only slightly over time. This suggests that consensus on the meaning and significance of new papers is achieved very rapidly. There is evidence, however, that the uniformity of usage is lower for these research front papers than for very highly cited method papers (76% versus 87%).

Using the *Cohen*'s *Scientific American* article we were able to place the cluster into close correspondence with a formal or official "history" of recombinant-DNA. Our purpose was not to show that either our 1976 cluster or *Cohen*'s article constituted a history of the specialty, but rather that both were versions of the paradigm for the field circa 1976. The relative locations of documents on the map were seen not to reflect a chronological progression but rather subject matter associations which reflected 1976 views on the relationships among the prominent concepts.

Citations context analysis enabled us to identify the concept associated with each highly cited document, and co-citation context analysis gave us a way to study the relationships which existed among the document-concepts. We performed

the latter analysis for only a few co-citation links (the bridges between regions), but if such an analysis were carried out for each of the links on the map, a complete cognitive structure for the research area could be obtained.

The technique of combining co-citation clustering and citation context analysis may provide a way of probing the structure of paradigms. We know that strong patterns of citation and co-citation exist and represent the choices of many different individuals. This suggests that the specialty community is behaving in a concerted way. There may be a sense in which different individuals are selecting and combining concepts in a coordinated way so that it appears that a collective "logic" or "rationality" is governing this behavior.

It seems likely that co-citation links can be rewritten in sentence-like form, and that entire single-link networks of the kind shown in Fig. 2 could be translated into patterns of interlocking sentences. What would be achieved then is a transformation of a bibliometric structure into a linguistic or word structure, thus rendering it amenable to logical analysis. If we regard the network of sentences as representing the instantaneous logical structure of a field, then we have opened a way of investigating the logic of scientific change at the specialty level. These possibilities await further research.

*

Work supported by National Science Foundation grant SOC76–08553.

References

1. M. MULKAY, *Science and The Sociology of Knowledge*, (London; Allen and Unwin, 1979).
2. H. G. SMALL, B. C. GRIFFITH, The Structure of Scientific Literatures I: Identifying and Graphing Specialties, *Science Studies*, 4 (October 1974) 17–40.
3. D. SULLIVAN, D. H. WHITE, E. J. BARBONI, State of a Science – Indicators in Speciality of Weak Interaction, *Social Studies of Science* 7 (May 1977) 167–200.
4. S. COLE, J. R. COLE, L. DIETRICH, Measuring the Cognitive State of Scientific Disciplines, in: *Toward a Metric of Science*, New York; Wiley and Sons, 1978, p. 209–251.
5. N. C. MULLINS, L. L. HARGENS, P. K. HECHT, E. L. KICK, Group Structure of Co-citation Clusters – Comparative Study, *American Sociological Review*, 42 (1977) 552–562.
6. M. J. MORAVCSIK, P. MURUGESAN, Some Results on the Function and Quality of Citations, *Social Studies of Science*, 5 (1975) 86–92.
7. D. E. CHUBIN, S. D. MOITRA, Content Analysis of References: Adjunct or Alternative to Citation Counting?, *Ibid.*, 423–41.
8. I. SPIEGEL-RÖSING, Science Studies: Bibliometric and Content Analysis, *Social Studies of Science*, 7 (1977) 97–113.
9. S. COLE, The Growth of Scientific Knowledge: Theories of Deviance as a Case Study, in: *The Idea of Social Structure*, L. A. COSER (Ed.), New York; Harcourt, Brace, Jovanovich, 1975, p. 175–220.

10. H. G. SMALL, Cited Documents as Concept Symbols, *Social Studies of Science*, 8 (1978) 327–340.
11. H. G. SMALL, A Co-citation Model of a Scientific Speciality: A Longitudinal Study of Collagen Research, *Social Studies of Science*, 7 (1977) 139–166.
12. P. H. A. SNEATH, R. R. SOKAL, *Numerical Taxonomy*, San Francisco; W. H. Freeman and Co., 1973, p. 131.
13. B. C. GRIFFITH, H. G. SMALL, The Structure of Scientific Literature II: The Macro- and Micro-Structure of Science, *Science Studies*, 4 (October 1974) 339–365.
14. J. A. HARITGAN, *Clustering Algorithms*, New York; Wiley and Sons, 1975, p. 199.
15. H. G. SMALL, Structural Dynamics of Scientific Literature, *International Classification*, 3 (1976) 67–74.
16. J. B. KRUSKAL, Multidimensional Scaling by Optimizing Goodness-of-Fit to a Non-metric Hypothesis, *Psychometrika*, 29 (1964) 1–37.
17. T. S. KUHN, *The Structure of Scientific Revolutions*, 2nd ed., Chicago; University of Chicago Press, 1970, p. 174.
18. S. N. COHEN, The Manipulation of Genes, *Scientific American*, 233 (1975) 25–33.
19. V. SGARAMELLA, Enzymatic Oligomerization of Bacteriophage P22 DNA and of Linear Simian Virus 40 DNA, *Proc. Nat. Acad. Sci.* U. S., 69 (1972) 3389–3393.
20. M. MANDEL, A. HIGA, Calcium-Dependent Bacteriophage DNA Infection, *J. Mol. Biol.*, 53 (1970) 159–162.
21. See S. N. COHEN, Gene Manipulation, *New England Journal of Medicine*, 294 (1976) 883–889, for an explicit reference to HERSHFIELD 74 in a very similar context.
22. See S. N. COHEN, *Ibid*, for an explicit reference to KEDES 75 in a very similar context.