Current Comments

Computer-Aided Historiography—How ISI Uses Cluster Tracking to Monitor the "Vital Signs" of Science

Number 14 April 5, 1982

On numerous occasions, I've said that ISI® 's various data bases will become increasingly more interesting to the historian of science.[1,2] However, there is a serious problem in using the expression "history of science." The term too often connotes scholarship confined to the ancient origins of science. Consequently, the subject matter of most history of science is thought to be so distant that it doesn't require the attention of the contemporary scientist.

But the progress of science has accelerated sharply in recent decades. Significant discoveries required decades or centuries for past generations of scientists to accomplish. Today, breakthroughs occur in months or years. This is manifested in an explosive output of literature. The number of papers published each year is orders of magnitude greater than in the past and, contrary to popular myth, the information content of the average scientific paper today is equal to or greater than the average paper published 50 years ago. We remember primarily the great papers of the past. We forget that many less dramatic papers were also published then.

For these and many other reasons, history should be a vital concern of every contemporary scientist and certainly of every research manager or administrator. While the original purpose of ISI's data bases was to simplify information retrieval, they have now become a working tool for the historian and sociologist of science.

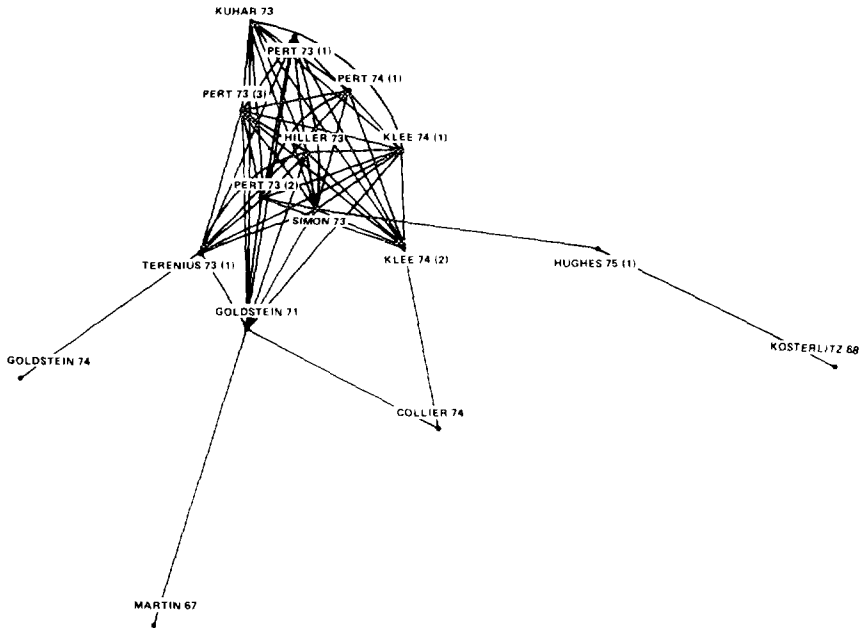
When we introduced the *ISI/BIOMED* ™ online system,[3] I described how our clustering techniques identify active research fronts. For each rapidly advancing field we identify the core papers, the principal investigators, and how closely or distantly related the fields are. This is graphically portrayed in the many individual "maps" we've published, which are highlighted in the encyclopedic *ISI Atlas of Science* ™.[4] Incidentally, the biochemistry/molecular biology section has now been published.[5]

A few years ago, we applied this technique to the field of opiate receptor research.[6] We identified the key people and core papers as that field developed year by year. For each year, we created a characteristic cluster map. In Figure 1, the map for 1975 is shown. In Figure 2, all of the separate annual maps of opiate receptor research have been consolidated and represented in a single historical map or flowchart of the field. This single composite "string" of maps shows how the field has changed and branched over the first six years.

This "cluster tracking" procedure is, in fact, computerized historiography. It is an extension of a simpler method for mapping science that was first demonstrated by Irv Sher, ISI's director of quality control, and me when we "wrote" the history of the genetic code using citation data.[7]

During the past year, I've shown these maps to many groups in the US and abroad. I've sensed the excitement in

Figure 1: 1975 cluster map: "Opiate Receptors."

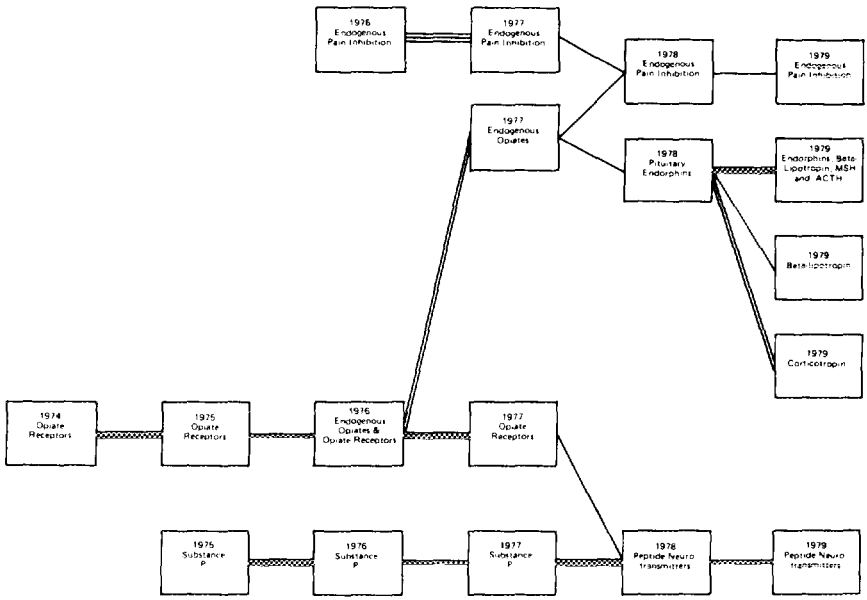


my audience whenever these annual cluster maps and the composite cluster strings as shown in Figure 2 are presented. This excitement increases when I point out that in the not too distant future, each user will be able to ask our online system to help "write" the history of any scientific field. Sometimes we call this *ad hoc* clustering. You can do this simply by selecting any starting point, like a milestone paper or a pair of highly co-cited papers. But at ISI we create clusters on a large scale using large computers to "batch process" the citation data in the *Science Citation Index®* (*SCI®*) file. We do this not only for the entire *SCI* file but also for various subdivisions. We are able to obtain clusters that identify more specific fields of information within each subdivision. From all this we create thousands of clusters each year.

Table 1 shows the number of clusters produced each year from 1970 to 1979, the size of the annual *SCI* files, and the citation and strength thresholds for each year. Citation and strength thresholds define whether or not a paper is eligible to appear in a cluster. The citation threshold sets the minimum number of citations a paper must receive in a given year to qualify for clustering. This threshold has increased from 15 to 17 citations as the *SCI* file has grown over the years. In addition, core papers must be cited *together* (co-cited) at a minimum "strength"—that is, the percentage of their total citations that are co-citations. Again, we have increased the minimum strength threshold from 16 to 22 percent as our files have grown. For a detailed discussion, refer to my earlier essays on the ABCs of cluster mapping,[8,9] reprints of which are available from ISI.

The purpose of this essay is to give you a general idea of how we go from the formation of a single cluster in a given year to a string of clusters span-

Figure 2: Pain and neurotransmitters cluster string: 1974-1979.



ning several years. Much of the work I'll describe was conducted by Roberts Coward, ISI's senior research projects associate, with support from the National Science Foundation.[10] To illustrate the cluster tracking procedure, I've selected clusters of papers on plant protoplast research—a field that has been quite active in the last 15 years.

Plant protoplasts are plant cells without their walls. The two basic components of protoplasts are the nucleus and the surrounding cytoplasm, which are bound by a membrane. Plant scientists are interested in fusing protoplasts of cells from different plant species to create hybrid cells. These hybrid cells would divide and multiply, and eventually differentiate into a whole new plant that combines characteristics of its "parents." The potential impact of this research, if successful, would be enormous. For example, new hybrid plants might be created which could thrive in harsh environments that do not now support plant growth, or new strains of disease-resistant plants might be developed.[11]

In the following exercise, we'll track the development of this field of research over several years. In order to track clusters, we developed a computer program that determines which of the several thousand clusters generated each year contain any of the core papers from a given cluster formed the previous year. The algorithm is simply a series of matching operations. That is, if we start with core papers in a 1973 cluster, the computer matches them with papers in the 1974 file. In fact, the algorithm also works *backward* in time—we could just as easily match the 1973 cluster with papers in the 1972 file. But for the sake of convenience, we'll track the plant protoplast clusters forward in time.

We'll start with a cluster we entitled "Culture and Use of Plant Protoplasts," shown in Figure 3. This cluster was generated from the 1973 *SCI* data base.

Table 1: Number of clusters generated from annual files of the *SCI*® data base, 1970-1979. "Size of file" refers to the total number of source items included in the data base. "Citation threshold" refers to the required minimum number of citations a *single* paper must receive in order for it to enter the clustering procedure. "Strength threshold" refers to the minimum percentage of co-citations between *two* papers' total citations required in order for the pair of papers to appear in a cluster.

Year	Size of File	Number of Clusters	Citation Threshold	Strength Threshold
1970	361,875	1199	15	16%
1971	364,490	1335	15	16%
1972	377,614	1382	15	16%
1973	406,943	1610	15	16%
1974	400,971	1702	16	16%
1975	418,903	1546	16	16%
1976	450,956	1928	16	18%
1977	494,861	2542	16	20%
1978	500,702	2350	17	22%
1979	517,557	2336	17	21%
Total	4,294,872	17,930		

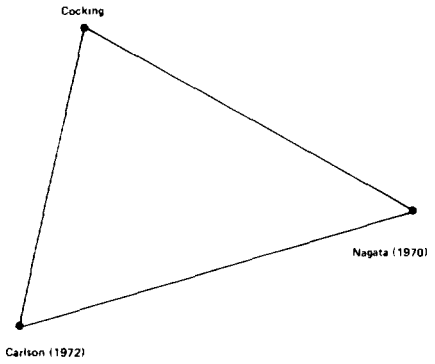
Each of the papers in this cluster was cited at least 15 times, and all were frequently co-cited by scientists publishing in 1973. In 1973, 35 papers cited one or more of the core papers. The cluster was named by examining rank-ordered lists of words that appeared in the titles of the citing papers published in 1973. While the naming of clusters is not completely automatic, the procedure is based on terminology in current use. So when you use it for retrieval purposes, you realize it is *not* based on any *a priori* list of terms in a thesaurus. The indexing is actually independent of the title words used in the retrieved papers. You can better appreciate this by using research front searching in one of ISI's data bases.

The full bibliographic information for the three core papers in this cluster is given in Table 2. The 1972 paper by Edward Cocking, University of Nottingham, England, provides a general review of research into the isolation, growth, and development of plant cell protoplasts. The 1970 paper by Toshiyuki Nagata and Itaru Takebe, Institute for Plant Virus Research, Chiba, Japan, reports that isolated protoplast cells divided after regenerating their walls. According to the authors, this is the first time that protoplast cell division was reported.[12] Finally, the 1972 paper by Peter Carlson and colleagues, Brookhaven National Laboratory, Upton, New York, reports the successful fusion between protoplasts from different plant species. In a sense, these three papers set the stage for future research into plant protoplasts—how they are isolated, how they develop, and how they can be fused.

Since this was an expanding field of research, it is not surprising that we identified a 1974 cluster which contained all of the core papers from the 1973 cluster. The 1974 cluster contains eight "new" core papers. These were in addition to the three from 1973, which are indicated by squares on the map in Figure 4. This 1974 cluster, shown in Figure 4, was also entitled "Culture and Use of Plant Protoplasts." The growth of this field is indicated by the number of citing papers. Whereas 35 papers cited the three core papers in the 1973 cluster, 110 papers cited the cluster of 1974 core papers.

Bibliographic information on the eight new papers appears in Table 2. Four of these papers are concerned with the same basic topics described in the core papers of the 1973 cluster—the isolation, growth, and fusion of plant protoplasts. The 1968 paper by Takebe and colleagues describes a method of isolating large numbers of plant protoplasts.

Figure 3: 1973 cluster map: "Culture and Use of Plant Protoplasts."



The 1971 paper by Nagata and Takebe describes an agar medium on which protoplasts can divide and grow. Another 1971 paper, by Takebe and colleagues, claims that protoplasts can grow into entire plants on this agar medium. The 1970 paper by J.B. Power and colleagues, University of Nottingham, deals with the fusion of plant protoplasts.

The remaining four papers introduce a topic that wasn't addressed in the 1973 core papers—the role of protoplasts in genetic research. The 1969 paper by Takebe and Yoshiaki Otsuki describes how viruses infect and multiply within plant protoplasts. The 1973 paper by Carlson points out the kinds of genetic questions that can be answered using plant protoplasts. The 1973 paper by C.H. Doy and colleagues, Australian National University, Canberra, concentrates on the transference of bacterial genes to plant cells. Finally, the 1971 paper by Carl Merril, National Institute of Mental Health, and colleagues deals with the introduction of bacterial genes into mammalian cells.

When the tracking process is continued, the 1975 file is searched for any of the 11 core papers from the 1974 cluster. As it turns out, due to the increased activity in this field, there are two 1975 clusters related to plant protoplast research. That is, some of the core papers in the 1974 cluster have continued over into two separate 1975 clusters. Both are shown in Figure 5.

The smaller cluster at the top of Figure 5, entitled "Protoplasts in Plant Virology," contains three core papers. Two of them are carry-overs from the 1974 cluster. These two papers, authored by Takebe, are indicated by squares on the map. The "new" paper is authored by Shigeji Aoki with Takebe as coauthor, and was published in 1969. All three papers concentrate on the infection of isolated protoplasts by tobacco mosaic virus. A total of 35 papers published in 1975 cited the three core papers in this cluster.

The larger cluster at the bottom of Figure 5, entitled "Culture and Use of Plant Protoplasts," includes seven core papers, six of which are carry-overs from the 1974 cluster. Again, these are indicated by squares. The 1974 paper by K.N. Kao and M.R. Michayluk, National Research Council, Saskatchewan, Canada, is the "new" paper. All seven papers in this 1975 cluster concentrate on the isolation and growth of plant protoplasts. A total of 92 papers cited the core publications in this cluster.

This tracking procedure is repeated by the computer, forward and backward in time, until no adjacent year links are found. As it turns out, the string of plant protoplast clusters extends from 1973 to 1980. Whether or not this string extends into 1981 will be determined when we finish clustering that year's file of *SCI* data. Figure 6 shows a flowchart of the entire string of plant protoplast clusters. Each box represents a cluster generated from the corresponding annual file of *SCI*. The box contains the year, the name, the number of cited core papers, and the number of citing papers published that year.

The lines between boxes indicate the "relatedness" or "stability" of clusters that are linked by core papers they

Table 2: Bibliography of papers appearing in Figures 3-5.

Aoki S & Takebe I. Infection of tobacco mesophyll protoplasts by tobacco mosaic virus ribonucleic acid. *Virology* 39:439-48, 1969.

Carlson P S. The use of protoplasts for genetic research. *Proc. Nat. Acad. Sci. US—Biol. Sci.* 70:598-602, 1973.

Carlson P S, Smith H H & Dearing R D. Parasexual interspecific plant hybridization. *Proc. Nat. Acad. Sci. US—Biol. Sci.* 69:2292-4, 1972.

Cocking E C. Plant cell protoplasts—isolation and development. *Annu. Rev. Plant Physiol.* 23:29-50, 1972.

Doy C H, Gresshoff P M & Rolfe B G. Biological and molecular evidence for transgenosis of genes from bacteria to plant cells. *Proc. Nat. Acad. Sci. US—Biol. Sci.* 70:723-6, 1973.

Kao K N & Michayluk M R. A method for high-frequency intergeneric fusion of plant protoplasts. *Planta* (Berl.) 115:355-67, 1974.

Merril C R, Geier M R & Petricciani J C. Bacterial virus gene expression in human cells. *Nature* 233:398-400, 1971.

Nagata T & Takebe I. Cell wall regeneration and cell division in isolated tobacco mesophyll protoplasts. *Planta* (Berl.) 92:301-8, 1970.

Nagata T & Takebe I. Plating of isolated tobacco mesophyll protoplasts on agar medium. *Planta* (Berl.) 99:12-20, 1971.

Power J B, Cummins S E & Cocking E C. Fusion of isolated plant protoplasts. *Nature* 225:1016-8, 1970.

Takebe I, Labib G & Melchers G. Regeneration of whole plants from isolated mesophyll protoplasts of tobacco. *Naturwissenschaften* 58:318-20, 1971.

Takebe I & Otsuki Y. Infection of tobacco mesophyll protoplasts by tobacco mosaic virus. *Proc. Nat. Acad. Sci. US—Biol. Sci.* 64:843-8, 1969.

Takebe I, Otsuki Y & Aoki S. Isolation of tobacco mesophyll cells in intact and active state. *Plant Cell Physiol.* 9:115-24, 1968.

share. Stability defines the proportion of cited papers that are common to both clusters. The formula used to calculate stability is as follows:

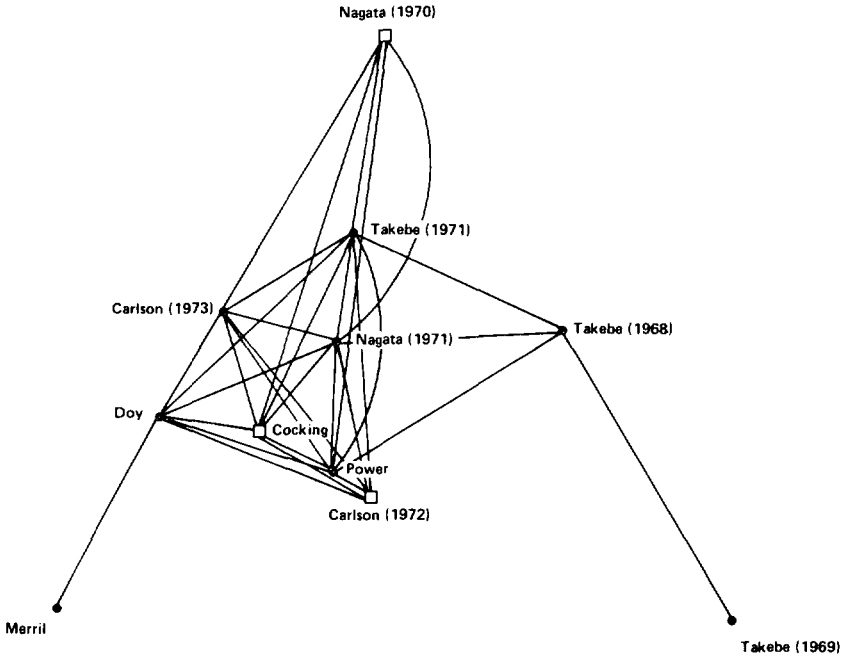
$N_{pA \rightarrow B} / (N_{pA} + N_{pB}) - N_{pA \rightarrow B}$,

where $N_{pA \rightarrow B}$ is the number of cited papers continuing from cluster A to B, N_{pA} is the number of cited papers in cluster A, and N_{pB} is the number of cited papers in cluster B.

For example, in Figure 6 we see that the 1976 cluster, "Plant and Mammalian Protoplast Fusion and Hybridization," includes three papers. The 1977 cluster to which it is linked, "Plant Protoplast Fusion and Hybridization," contains ten papers—all three of the 1976 core papers are included in this 1977 cluster. Thus, the stability of these two linked clusters is: $3/(3+10)-3=.3$, or 30 percent. In Figure 6, three lines are used to show that the link was formed at 30 percent or greater stability; two lines indicate stability between 20 and 29.9 percent; a single line indicates stability between ten and 19.9 percent.

These stability coefficients can be used as a rough indicator of how fast or slowly a field of research is developing. For example, let's consider two linked clusters that contain ten core papers each. In the extreme case of stability, all ten papers carry over from one cluster to the other, and there are no "new" core papers. This gives a stability value of $10/(10+10)-10=100$ percent. In the extreme case of "volatility," only one paper is carried over, and there are nine "new" papers. The stability value here is $1/(10+10)-1=5.3$ percent. However, stability is also affected by the number of core papers that are drawn into a cluster in the first place—this depends on citation frequency and co-citation strength thresholds set each year. As you can see in Table 1, these thresholds have varied over the years, and this may have an important impact on stability.

Figure 4: 1974 cluster map: "Culture and Use of Plant Protoplasts."

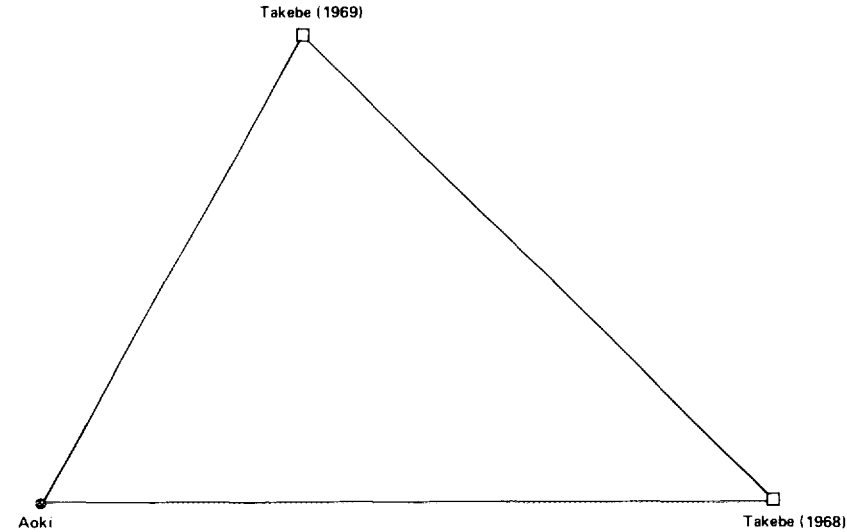


At ISI, we use stability coefficients to vary the number and size of cluster strings produced from the file of 17,930 clusters that have been generated from the 1970-1979 *SCI* data base (see Table 1). In order to do this, we set stability thresholds at four arbitrary levels—three, ten, 20, and 30 percent—and process the entire file of 1970-1979 *SCI* clusters. The idea is to find out which level produces strings of a manageable size. In general, increasing the stability threshold results in simplifying the average cluster string and increasing the total number of strings generated. These techniques are important in developing the cross-references we provide in our annual indexes to research fronts in *ISI/BIOMED* and *ISI/CompuMath*™.[13]

For example, at three percent stability, a total of 2,141 strings were formed with an average of 6.0 clusters per string. At ten percent, 2,803 strings were generated, containing an average of 4.3 clusters per string. More than 3,000 strings were formed at both 20 and 30 percent stability, but they averaged 3.3 and 2.8 clusters per string, respectively. These were too few to be interesting for analysis. Thus, we concentrated only on those cluster strings formed at three and ten percent. Table 3 shows complete statistics on cluster strings formed at these stability thresholds.

In Table 4 you can see the number of strings formed at three and ten percent stability that stretch over two to ten continuous years. The information in that table is graphically depicted in Figure 7. It's interesting to note that increasing the stability doesn't have a significant effect on the *proportion* of strings falling under each year span. For example, at three percent stability, 50.1

Figure 5: 1975 cluster maps: "Protoplasts in Plant Virology" (top) and "Culture and Use of Plant Protoplasts" (bottom).



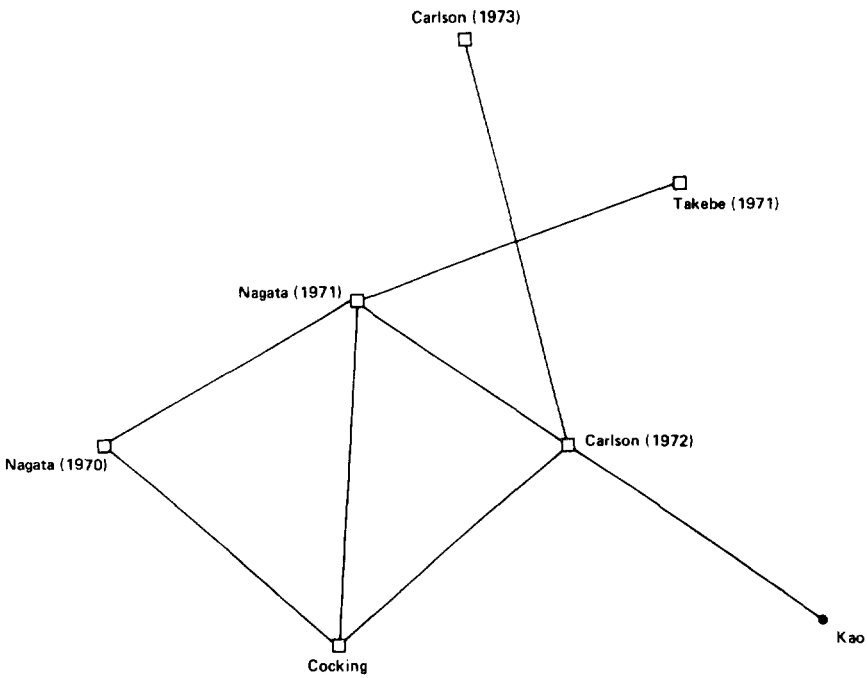
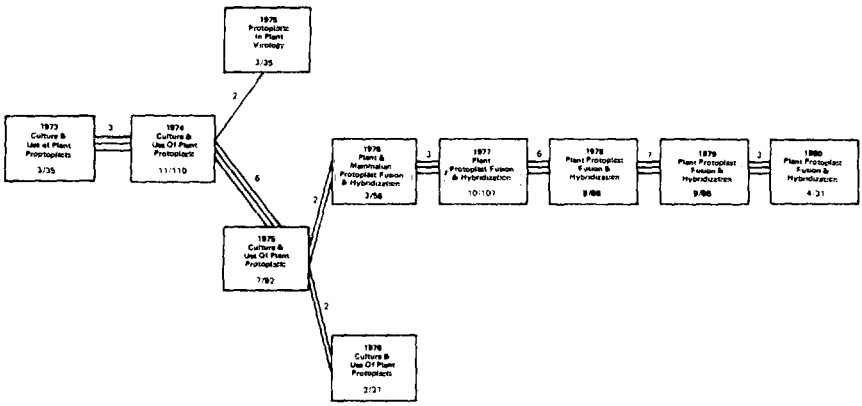


Figure 6: Plant protoplast cluster string: 1973-1980. Numbers at the bottom of each box refer to the number of cited/citing papers for each cluster. Numbers above the lines between boxes refer to the number of cited papers that carry over from one cluster to another. Lines between boxes indicate stability of linked clusters: three lines signify stability greater than or equal to 30 percent; two lines signify stability between 20 and 29.9 percent; one line signifies stability between ten and 19.9 percent.



percent of the 2,141 cluster strings span two years and 3.4 percent span all ten years. At ten percent, these figures are nearly identical—47.3 percent of the 2,803 cluster strings span two years and 3.6 span ten years.

Note that the distribution of cluster strings in Table 4 is almost exponential. That is, if you increase the span by a single year, the number of cluster strings is reduced to half; an increase of two years reduces the number of strings to a quarter; an increase of three years reduces the number to a ninth, and so on. Thus, it becomes increasingly rare for cluster strings to span two, three, four, or more years. This may indicate that modern scientific research is very volatile—only a very small number of papers remain core to their field for more than a few years. However, this conclusion may not apply to the many smaller research fronts that are not identified at the levels we are using.

But even for fast-moving research this conclusion must be interpreted cautiously. While few papers continue over five or more adjacent years, many may still appear in *nonadjacent* years. In other words, cited papers in a 1974 cluster that don't appear in 1973 or 1975 clusters may "resurface" in 1972 or 1976 clusters. Although we are interested in developing a more sophisticated computer program that identifies cluster links between nonadjacent years, it is more important that we verify that the historical maps we can already create from cluster strings correspond to the perceptions of scientists working in those fields.

At this point, cluster tracking enables us to view the histories of various lines of research. We can actually see how strings split or merge, and continue or terminate, as shown in Figures 2 and 6. When strings split, this may be a sign that the specialty is shifting onto a new plane because new discoveries are changing the focus of research. When strings merge, as in Figure 2 where opiate receptor research merged with the string for

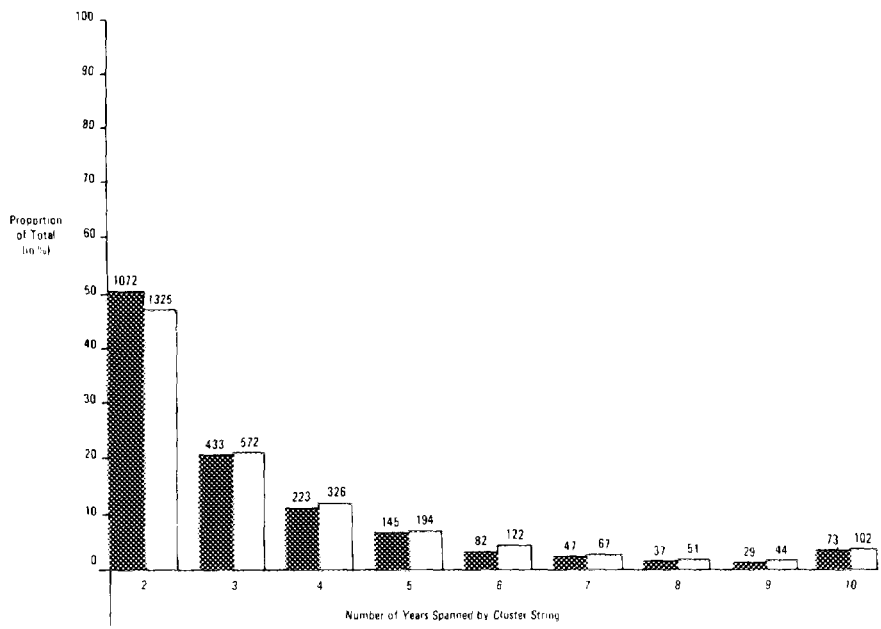
Table 3: Statistics for cluster strings formed at three percent and ten percent stability, *SCI*® data base, 1970-1979. Average string density is calculated by dividing the total number of clusters in strings by the total number of strings formed. "Simple strings" are linear strings—that is, they continue in a single line and do not split or merge.

Stability Threshold	Total Number of Strings Formed	Total Number of Clusters in Strings	Average String Density	Total Number of Simple Strings
3	2141	12,814	6.0	1719
10	2803	12,071	4.3	2139

Table 4: Year span distribution of cluster strings formed at three percent and ten percent stability, *SCI*® data base, 1970-1979.

Three Percent Stability			Ten Percent Stability		
Year Span	Number of Strings	Proportion of Total	Year Span	Number of Strings	Proportion of Total
2	1072	50.1%	2	1325	47.3%
3	433	20.2%	3	572	20.4%
4	223	10.4%	4	326	11.6%
5	145	6.8%	5	194	6.9%
6	82	3.8%	6	122	4.4%
7	47	2.2%	7	67	2.4%
8	37	1.7%	8	51	1.8%
9	29	1.4%	9	44	1.6%
10	73	3.4%	10	102	3.6%
Total	2141	100.0%	Total	2803	100.0%

Figure 7: Relative year span distribution of cluster strings formed at three percent and ten percent stability, *SCI*® data base, 1970-1979. Black bar indicates three percent stability. White bar indicates ten percent stability. Absolute numbers of cluster strings appear at top of bars.



Substance P, an important new consensus may be emerging regarding the nature of peptide neurotransmitters.

Thus, we can monitor a variety of the "vital signs" of science by examining the scientific literature through the technique of cluster tracking. Undoubtedly, many refinements will be needed to make the system more precise. In the hands of the informed expert who can provide appropriate cognitive insights, computer-aided historical research may help bridge the gap that often separates historians of science from those doing current research.

It would be absurd to assert that there is no difference between the scholarship of the history of "old-fashioned" science and that involved in doing the history of contemporary science. The historian writing about science during the tenth or eleventh century must not only deal with various linguistic problems but must also work with rare books or manuscripts. The modern historian has a wealth of review literature, correspondence files, recorded interviews, and biographical accounts to absorb. Hopefully, citation analysis and cluster tracking will benefit both the contemporary historian and the classical historian. To ease their tasks, ISI is presently completing citation indexes for 1955-1960—it's part of a program that will eventually cover the literature of the twentieth century.

* * * * *

My thanks to Patricia Heller and Alfred Welljams-Dorof for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Citation analysis as a method of historical research into science. *Citation indexing—its theory and application in science, technology, and humanities.* New York: Wiley, 1979. p. 81-97.
2. ----------. Scientometrics comes of age. *Essays of an information scientist.* Philadelphia: ISI Press, 1981. Vol. 4. p. 313-8.
3. ----------. ISI's on-line system makes searching so easy even a scientist can do it: introducing METADEX automatic indexing & ISI/BIOMED SEARCH. *Current Contents* (4):5-8, 26 January 1981.
4. ----------. Introducing the *ISI Atlas of Science: Biochemistry and Molecular Biology, 1978/80. Current Contents* (42):5-13, 19 October 1981.
5. **Institute for Scientific Information.** *ISI Atlas of Science™ : Biochemistry and Molecular Biology, 1978/80.* Philadelphia: ISI, 1981. 540 p.
6. **Garfield E.** Controversies over opiate receptor research typify problems facing awards committees. *Essays of an information scientist.* Philadelphia: ISI Press, 1981. Vol. 4. p. 141-55.
7. **Garfield E, Sher I H & Torpie R J.** *The use of citation data in writing the history of science.* Philadelphia: ISI, 1964. 86 p.
8. **Garfield E.** ABCs of cluster mapping. Part 1. Most active fields in the life sciences in 1978. *Essays of an information scientist.* Philadelphia: ISI Press, 1981. Vol. 4. p. 634-41.
9. ----------. ABCs of cluster mapping. Part 2. Most active fields in the physical sciences in 1978. *Essays of an information scientist.* Philadelphia: ISI Press, 1981. Vol. 4. p. 642-9.
10. **Coward H R.** *Tracking scientific specialties: indicator applications of time series co-citation clusters.* Unpublished paper, 1980. 19 p.
11. **Pontecorvo G.** Genetics, somatic cell. *McGraw-Hill encyclopedia of science and technology.* New York: McGraw-Hill, 1977. Vol. 6. p. 120A-120C.
12. **Nagata T & Takebe I.** Cell wall regeneration and cell division in isolated tobacco mesophyll protoplasts. *Planta* (Berl.) 92:301-8, 1970.
13. **Garfield E.** *ISI/CompuMath,* multidisciplinary coverage of applied and pure mathematics, statistics, and computer science, in print and/or online—take your pick! *Current Contents* (10):5-10, 8 March 1982.