Complementary Structures in Disjoint Science Literatures

Don R. Swanson
Center for Information and Language Studies
University of Chicago

Introduction

Difficult and intriguing information retrieval (IR) problems derive from what I call complementary but disjoint (CBD) structures within the literature of science. *Complementary* refers to the relationship between two separate scientific arguments which, when combined, yield important inferences and insights not apparent in the separate arguments. Corresponding to the two arguments are two *complementary literatures*. Each literature (ideally) is the "complete" set of articles that contain the argument in question. *Disjoint literatures* have no articles in common, do not cite or mention each other, and are not co-cited. If two complementary literatures are also disjoint, the possibility is worth investigating that the combined arguments and the inferences to which they lead might not be made explicit anywhere within the published record of science. The ever-increasing fragmentation of science into mutually-isolated specialties probably assures a limitless supply and combinatorial growth of implicit connections, some of which may be unknown solutions to important problems. These solutions are worth seeking.

I have previously analyzed and reported three examples of CBD literature pairs, each of which led to a novel and plausible medical hypothesis that appeared to merit testing [1-5]. The universe of published medical literature, not a selected subset, was the target of a process that depended not only on extensive use of online information services but also on reading the text of many hundreds of medical articles.

My claim that novel inferences can be drawn from CBD literatures requires clarification. CBD literatures of course can never attain their ideal state of completeness as specified in the definition, but they are presumed to approach as near to completion as attainable with a diligent search of the literature. On that basis, one can stake a claim to novelty in the usual way, by publishing it. The domain of the claim is then the universe of scientific knowledge, within which the claim is open to criticism and rebuttal. Two studies that I published [1,2] were followed one to three years later by independent clinical and laboratory tests that supported the corresponding hypotheses that I had proposed [6-8]. Only one of the three papers cited my work [2, 7], but, more important, no work prior to mine concerning the connections in question was cited. The publication sequence thus provides evidence to support my central claim that, in principle, new hypotheses can emerge, and scientific discovery can be anticipated or stimulated, through the investigation of CBD literatures.

SIGIR91: Proceedings of the 14th Annual International ACM SIGIR Conference; Oct 1991, Chicago. NY: ACM Press

Open vs. Closed Domains

The difficulties inherent in conducting research within a universal domain as defined above suggest that more limited objectives merit consideration. One might develop a methodology for discovering novel connections within a closed and highly incomplete domain, for example.*

One potentially fatal shortcoming of such an approach is that novelty then might depend more on the defects of the domain than on the merits of the method. Any synthesis of separate ideas or arguments can be made novel with respect to a sufficiently restricted domain — for example a domain formed by removing all cross-citing and co-citing articles from complementary literatures in order to create a disjoint subset of the same literatures. A methodology demonstrated on such an artifactual basis would be difficult to evaluate.

Notwithstanding the variability and unreliability of individual citation practices, disjointness in the open domain of science can be highly reliable for sufficiently large literatures. Even if the probability is as low as 3% that any particular relevant article will be cited by a given author, the compound probability would be less than 5% that 100 independent failures would occur by chance across a pair of disjoint literatures. If, on the average, five articles in

*I am grateful to Abe Bookstein for suggesting the key idea that novelty is necessarily domain dependent.

one literature are relevant to each article in the other, then 100 citation failures would occur in disjoint literatures that consist of 20 articles each. In the examples I have investigated so far, the individual CBD literatures have ranged from 30-60 articles.

To embed research on CBD structures within the open domain of science, as I have chosen to do, has the overriding practical advantage of forcing a more intimate confrontation with problems of scale. It does not take more than a few thousand Medline searches to become convinced that IR problems on the scale of six million records (300,000 on neoplasms alone) are qualitatively different from those we encounter in small-scale experiments. I shall try to make clear at a later point how some of these differences become manifest.

My main purpose in this paper is to examine in detail two complementary structures, the first based on an already existing historical example and the second based on one of the studies previously reported.

Complementarity in the Sutton Synthesis

The Sutton-Boveri hypothesis of 1903 on the connection between genes and chromosomes showed how important scientific insights can be derived by a synthesis of ideas already in the literature [9]. Sutton's paper has been described by various authors as a combination of ideas from two areas of investigation, a confluence of two streams of effort, a merger of two independent disciplines, and as marking the end of the separate histories of cytology and genetics and the beginning of cytogenetics. Darden and Maull described the Sutton synthesis in terms of what they called an interfield theory, and they show how such theories can lead to new predictions in each of the two component fields [10].

One of the fields (Sutton's own specialty) was represented by hundreds of articles on chromosomes, dating back to 1876. The other, to judge only by what Sutton cited, was represented by a single paper — the famous long-neglected 1866 paper on pea hybridization by Mendel that surfaced in 1900 [11]. What is of interest for our purposes here is the relationship of complementarity between the chromosome literature and Mendel's paper.

The salient features of the chromosome literature lie in the two processes of meiosis and fertilization. Meiosis includes a reducing division in which paired homologous chromosomes separate to form gametes with half as many chromosomes as are in the somatic cell nuclei. Fertilization consists essentially in the conjugation of two gametes to form a zygote in which the normal diploid count of chromosomes is restored. The chromosome literature before 1903 contains voluminous descriptions in minute detail, accompanied by thousands of sketches, of the cell nucleus as seen through a microscope. It was generally accepted well before Sutton that chromosomes must be the material basis for the transmission of hereditary traits.

Mendel's paper describes what peas look like. Seeds, for example, can be round or wrinkled, yellow or green. Mendel sought to discover the laws of formation and development of hybrids. He found that crossings between two plants differing in a single pair of traits (A,a) produced a hybrid that looked just like one of the parents; that hybrid exhibited what Mendel called the dominant trait, A. The next generation, breeding the hybrid group alone, yielded two forms that occurred in a ratio of 3:1, dominant to recessive, each form resembling one of the two grandparents. The recessive form was then found to breed true, while the progeny of the dominant form split into two groups in a ratio of 1:2; the first bred as pure dominant, and the second bred just like the hybrid. The 3:1 ratio thus could be seen as actually representing three groups that occurred in a ratio of 1:2:1, or A+2Aa+a. One would expect just such a result if the two traits A,a split up within the new seeds and pollen, and then recombined at random during fertilization. Mendel went on to show that a second pair of traits B,b behaved in exactly the same way and as though entirely independent of the first pair. He concluded that the progeny of hybrids can be represented by the terms of a combination series in which the series for each pair of differing traits are independently combined. Thus the series for two traits can be derived from all combinations that occur in (A+2Aa+a)(B+2Bb+b).

What makes the above two paragraphs complementary is their common implication of a combinatorial process. The two homologous chromosomes that pair up are presumed to have originated from the two parents. As they move onto the equatorial plane of the spindle, prior to dividing, Sutton noted that they are just as likely to be oriented toward one pole as the other. The gametes that are formed then are just as likely to receive any given chromosome from one parent as the other. Sutton saw that if he associated a pair of parental traits A,a with a pair of parental chromosomes, then he could account for Mendel's observation that traits appeared to split up in the germ cells; as many A-type gametes as a-type would be formed. Upon subsequent fertilization, random encounters between gametes would yield a zygote distribution that can be written as A+2Aa+a. A similar argument could be made if a second pair of traits were identified with a second pair of chromosomes. Moreover, it would behave independently of the first pair, thus accounting for Mendel's laws of segregation and distribution. This was the essence of the complementarity that Sutton saw, and on the basis of which he postulated the association between genes and chromosomes.

Thus a problem or question posed in one of two disjoint literatures — the problem of how a pair of traits can behave as though they were randomly distributed to the progeny — is resolved implicitly in the other; the processes of meiosis and fertilization provided a causal mechanism sufficient to explain Mendel's results. Here complementarity bridges a gap between the microlevel of the cell nucleus

and the macrolevel of visible traits in mature plants. The languages of description and argument in the two areas are quite dissimilar.

This example of complementarity poses a formidable challenge to workers in the vineyards of artificial intelligence, knowledge representation, and language processing. The heart of the problem of automating the discovery of such complementarity is to provide a computer with the kind of knowledge and creativity that enabled Sutton to see the crucial question — the orientation of the coupled pair of chromosomes at the moment of separation in meiosis.

Prospective vs. Retrospective Complementarity

Using the Sutton synthesis as a point of departure, I am presently investigating the possibility that the cytology and hybridity literatures long before 1903 qualified as CBD. Looking back to the 19th century, knowing about the Sutton synthesis, one can focus immediately on the relevant literatures. A prospective study, on the other hand, must begin with only half of the argument, or rather with the statement of some interesting problem, without knowing where the complementary half — the solution — is to be found.

I examine next the logical structure of one of the three examples of CBD literatures. I shall not attempt to describe the actual process by which these literatures were found, but instead will focus on the end result and its implications for designing a computer-aided process for arriving at that result. The example chosen is based on the medical problem of finding a cure for migraine headache [2].

Migraine and Magnesium: Eleven Complementary Connections

Migraine headache is a mysterious disorder for which neither cause nor cure is known. The original problem, starting with the literature on migraine (the "source" literature), was to find somehow a second or "target" literature, not about migraine, that leads one to see a possible unreported cause or cure. The problem of how to conduct such an exploratory search without knowing the destination in advance is discussed elsewhere [2, 12, 13]. Here I assume that the target literature (magnesium) has already been identified; my purpose is to analyze the relevant logical connections. A subset of that literature showed that dietary magnesium can lead to certain specific physiological changes. At the same time, a subset of the migraine literature showed that the same kind of changes might benefit patients with migraine headache. Yet, with a few exceptions (discussed in [2]), these two complementary literatures were disjoint. Neither literature mentioned the other or suggested what they jointly imply — that magnesium deficiency might be a causal factor in migraine.

Eleven indirect connections in these CBD literatures are summarized in Table 1. Each statement labelled "a" is about migraine and the corresponding statement "b" is about magnesium; the italicized phrases shown are common to the two statements and form the connecting link. There are eleven pairs of complementary "a—b" statements; presumably (judging from the literature) no one person is acquainted with both members of any pair. The subset of migraine literature on which the "a" statements are based consists of 65 articles, and the magnesium subset consists of 63 articles that lead to the "b" statements. These 128 articles, and others, have been reviewed in depth in order to construct the complete argument, which has been published in a biomedical journal [2]. It is worth stressing that Table 1 consists of a constructed *summary*, not of extracted text from the migraine and magnesium literatures.

TABLE 1: Logical Structure of Implicit Connections between Migraine and Magnesium Literatures

Notation for a "logic of suggestibility":

—> means "can or might influence" (e.g., the course of a disease — such as either a cause, cure, amelioration, exacerbation, etc.) Influence may be indirect, via any number of links, and there is no presumption that the influence is exclusive — that is, there may well be other influences. It is intended to signify a plausible conjecture.

Thus A —> L and L —> M implies (conjecturally) A —> M
But also, less obviously, A —> L and A —> M might imply, conjecturally, L —> M or M —> L since either L or M might be an intermediate link.

"==" means "equivalent in action" or "equivalent in mechanism" depending on whether it connects two drugs or two diseases. Or, it can be taken to mean "is comparable to".

Let A = magnesium — either dietary or internal to the body.
Implicitly refers, in most cases, to quantity of magnesium.
Let M = migraine headache
Let L = one or more intermediate physiological links
Ln = "n"th link if there is more than one to be denoted

---

1. a) *Stress and Type A behavior* are associated with migraine.
   b) *Stress and Type A behavior* lead to body loss of magnesium.
   Let B = Stress and Type A behavior
   a) suggests that B might lead (even if indirectly) to M. That is, B —> M
   b) implies that B can influence A. B —> A
   Thus A might be one of the intermediate links, and so can conclude that A might influence M. A —> M

2. a) Excessive *vascular tone and reactivity* may increase susceptibility to migraine.
   b) Magnesium can reduce *vascular tone and reactivity.*
   Let L = vascular reactivity.
   a) suggests L —> M
   b) suggests A —> L Conclusion: A —> M

3. a) *Calcium channel blockers* have been used successfully in preventing migraine attacks.
   b) Magnesium is a natural *calcium channel blocker.*
   Let B = calcium channel blockers

a) suggests B —> M
b) suggests A == B Thus A —> M

4. a) "*Spreading cortical depression*" is thought to be implicated in the early phase of a migraine attack.
b) High levels of magnesium in the extracellular cerebral fluid inhibit *spreading cortical depression* in laboratory animals.
Let B = Spreading cortical depression
a) B == M b) A —> B Thus A —> M

5. a) There is evidence for a connection between *epilepsy* and migraine.
b) Magnesium deficiency may increase susceptibility to *epilepsy*.
Same as 4, but with B = "epilepsy" instead of "spreading depression".

6. a) Migraine patients have abnormally high *platelet aggregability*.
b) Magnesium can suppress *platelet aggregation*.
L = platelet aggregation
Same as 2 except for definition of L.

7. a) Platelets of migraine patients are abnormally sensitive to *serotonin release*.
b) Magnesium can inhibit *serotonin-induced* contractions of vascular smooth muscle.
L = serotonin release by blood platelets
L2 = vascular contraction
a) L —> M
b) L —> L2 Thus L2 may be intermediate, so L2 —> M
b) also implies A —> L2
Thus A —> M

8. a) *Substance P* may be a cause of head pain in migraine.
b) Magnesium has an antagonistic effect on *Substance P* activity.
L = Substance P, otherwise same as 2 and 6

9. a) Abnormal *prostaglandin (PG) release* can aggravate vasoactivity in migraine.
b) Magnesium increases *prostacyclin (PGI2) formation*.
L = prostaglandin
L2 = vasoactivity
a) L —> L2 —> M and so L —> M
b) A —> L
Thus A —> M

10. a) Migraine may involve sterile *inflammation* of the cerebral blood vessels.
b) Magnesium has anti-*inflammatory* properties.
L = inflammation
a) L == M or L —> M
b) A —> L Thus A —> M

11. a) Cerebral *hypoxia* may play a key role in migraine.
b) Magnesium can protect against brain damage from *hypoxia*.
L = cerebral hypoxia, otherwise same as 2, 6, and 8.

The summarized arguments in Table 1 reveal a number of prototype logical structures. The simplest structure, which we may take as paradigmatic, is a causal syllogism: A —> L; L —> M; thus A —> M. Statements 2ab in Table 1 have essentially this form, where A represents magnesium, L represents vascular tone and reactivity, and M represents migraine. Statement pairs 6, 8, and 11 have the same structure as 2. The arrows are to be read as "can or might influence". We need not assume that the symbolized relationships are necessarily transitive, for the question of interest here is whether they are "suggestive" — that is, whether a person confronted with the two arguments that A —> L and L —> M would immediately guess that A might influence M. In general, very few AL-type or LM-type statements of interest are likely to be conclusively established or free of dispute, nor are most influences likely to represent sole or unique causes. Any AM-type conclusion no matter how plausible is always to some degree conjectural and would have to be put to a direct test.

Moreover, suggestive structures more complex than our paradigmatic syllogism also occur frequently. Each of the eleven pairs of questions is analyzed symbolically in Table 1, in order to exemplify some of these structures. There are doubtless many more. The fact that as many as eleven parallel links were found in this case lends a certain robustness [14] to the common inference to which they all lead, A —> M, even though some of the individual links may be tenuous. Unlike gene-chromosome complementarity, these links do not bridge a gap between two levels of observation, nor do they involve dissimilar vocabularies.

In sum, we are interested in what might loosely be called a "logic of suggestibility", the aim of which is to stimulate new and plausible hypotheses. Such a logic appears to have the properties that Wimsatt (citing Lenat and McCauley) has identified as characteristic of a heuristic procedure [15, p. 295].

The relationships between the logical form of the summary statements in Table 1 and the published text passages that warrant such statements are complex and elusive, but a few relationships appear to be simple and indeed can be seen within some of the titles alone. For example, the following six pairs of titles, taken from the reference list in [2], can readily be associated with statements 3 through 8 of Table 1, respectively, and clearly suggest to the reader a migraine-magnesium connection:

3a Role of *calcium entry blockers* in the prophylaxis of migraine.
3b Magnesium: nature's physiologic *calcium blocker*.

4a Leao's *spreading depression*: evidence supporting a role in the migraine aura.
4b Low extracellular magnesium induces epileptiform activity and *spreading depression* in rat hippocampal slices.

5a The relation of migraine and *epilepsy*.
5b Preliminary report on the magnesium deficient rat as a model of *epilepsy*.

6a Evidence of enhanced *platelet aggregation* and platelet sensitivity in migraine patients.
6b Protective effects of dietary calcium and magnesium on *platelet function* and atherosclerosis in rabbits fed saturated fat.

7a *Serotonin*-releasing factors in migrainous patients.

7b The effect of magnesium on the response of smooth muscle to 5-hydroxytryptamine [*serotonin*].

8a *Substance P* and enkephalins: a creditable tandem in the pathophysiology of cluster headache and migraine.

8b *Substance P*, bethanechol, 4-aminopyridine, and potassium antagonize the depressing effects of low-frequency stimulation, tetrodotoxin, procaine, and of magnesium on the field-stimulated guinea-pig vas deferens.

Again we may note a sharp contrast with the complementary structures in the Sutton paper. The connections here, as reflected in the italicized phrases, are easily recognized, and would seem to be far more tractable for approaches based on language processing, text "understanding", and IR probabilistic techniques.

With respect to "text understanding" systems, a few comments concerning background knowledge may be pertinent; the apparent ease with which one can infer from these titles that magnesium might influence migraine is somewhat deceptive. Neither the syntactic nor the semantic structures offer a straightforward chain of deductive reasoning with respect to a possible causal connection. The inference about a connection depends subtly on background knowledge that one brings to bear in understanding each title. In title 5b, for example, the word "model" carries a good deal of freight not reflected directly in its definition. (In this context, it is clear to a medical researcher that magnesium deficiency is what caused the epilepsy; is that how "model" should always, or usually, be understood?) In title 6a the idea of enhanced platelet aggregation must be linked with the allusion in 6b to a protective effect on platelet *function*, but that linkage depends on understanding "enhanced" as "excessive", what "platelet function" entails, and why there is some sort of a need for protection. To understand why 7a and 7b are brought together, one must understand the connection between ~~migraine~~ magnesium and smooth muscle. In title 8a, the phrase "creditable tandem" is literally nonsensical though metaphorically quite clear.

To appreciate the importance of structures other than the syllogisms mentioned earlier, it is instructive to examine titles 4a and 5a. The idea that there might be a connection between migraine and spreading depression or between migraine and epilepsy is difficult to express formally. Extensive and interesting but fuzzy arguments are offered in the literature. Each of these three phenomena is complex, and probably entails a cascade of events that constitutes what is usually called the "mechanism" of the disorder. The nature of any connection between two such phenomena is perhaps best expressed by saying that they may have in common certain stages of their respective mechanistic cascades. If so, a substance that either causes or cures one disorder, and that happens to act on a crucial common stage, might then have a similar effect on the other disorder. For purposes of symbolizing suggestibility, then, we define a loose sort of equivalence relationship between the two disorders (denoted "==" in Table 1).

So far as going much beyond what can be seen in Table 1 with respect to formalizing natural language text, I am more impressed with the problems than the prospects. We bring to bear an impressive bundle of presuppositions and background knowledge of human physiology, scientific method, scientific rhetoric, and metaphor, usually without being aware of it, in order to understand text. The study of CBD literatures may contribute substantially to AI's reservoir of insurmountable opportunities.

Prospects for a fully automatic search process

The six title-pairs listed above suggest a problem for which the computer would appear ideally suited. Automatic indexing and retrieval depends on representing documents or text passages by the words they contain; the similarity between two passages, or between a query and a document, is then in part a function of how many words they have in common, together with the certain information on relative frequencies. These titles suggest that, although the words "migraine" and "magnesium" do not appear together in any title, they are strongly linked by third parties that provide a shared context — namely "calcium blockers", "spreading depression", "epilepsy", "platelet", "serotonin", and "substance P". Searching for everything on migraine, we would therefore find an unexpected link to magnesium, and without knowing in advance that we were looking for magnesium! If we embed these titles in a relatively limited context of a few dozen or a few hundred other titles, one could no doubt dramatically demonstrate automatic discovery of the migraine-magnesium link. However, the picture changes altogether if we consider the actual context of these titles.

As of June, 1989, there were 2900 records in Medline with "migraine" in the title, 6500 with "magnesium", 12000 with "serotonin", 24000 with "platelet", 3000 with "calcium blockers", 13000 with "epilepsy", 360 with "spreading depression", and 3000 with "Substance P". The number of articles in the six intersections between these last six sets and the migraine set are, in the order listed: 103, 115, 42, 49, 8, 2. The corresponding numbers for "magnesium" are: 16, 24, 9, 26, 4, 1. The shared context is still there, but it is deeply buried. Only very small percentages of the literatures are linked. Yet it is such linkage that we must exploit in any fully automatic approach to retrieval. A preliminary study of this problem did not suggest that a fully automatic process would be able to distinguish "magnesium" from dozens or even hundreds of its uninteresting companions similarly distributed.

We turn instead to methods that involve some degree of intelligent human participation in the search process.

Co-occurrence Analysis of Title Words and Phrases

We need not abandon the idea of using title word co-occurrence data as an aid to navigation [16] in a partially

automated process. Beginning again, we consider the problem of finding the cause of migraine, or a cure for it. We proceed then on an unknown path toward an unknown destination. Guided by the example of six pairs of titles above, we first obtain a list of words that co-occur with "migraine", then identify all titles that contain these co-occurring words without "migraine". Next we identify the words that co-occur, in titles, with the words in the first list (noting from the example that magnesium is presumably among them).

To obtain such data, we downloaded (from Medline) 1000 titles containing the word "migraine". Physiologically significant key words (other than "migraine"), short phrases, and meaningful word combinations, were then selected, in a manual process, from each title. All such words or phrases that alluded to a process or to a substance that conceivably could be implicated in a causal chain of events leading to migraine were selected. The selection process essentially involved excluding obviously unsuitable words.

To give some idea of the kinds of words excluded, the following two nonexistent titles were constructed from pieces of several titles:

The prophylaxis of migraine: two double-blind comparative studies in more than 400 patients.

Clinical and experimental evidence for a role of drugs in the treatment of migraine; alterations related to the phase of the attack.

None of the words in either title appeared to offer any reasonable hope of pointing to causal connections. All would be rejected in the process described. ("Drugs" is too broad).

The selected words and phrases were organized into approximately 160 search terms by bringing together and logically combining related terms, including synonyms and inflectional variants. (Applying hindsight, we can notice that, among the 160 terms, 22 could be identified with 8 of the 11 intermediate literatures that were known to connect migraine to magnesium as indicated in Table 1.) The ratio of frequency of occurrence within migraine titles to frequency in all titles in the Medline database was computed for each term and used as a basis for ranking the 160 terms.

84 terms had a ratio greater than 8 — that is, each of these 84 terms occurred within migraine titles with a frequency 8 times greater than would result from a random distribution of such terms among all Medline titles. A cutoff of 8 was chosen somewhat arbitrarily but based on the (post hoc) argument that the 84 highest-ranking terms included 17 terms (out of the above-mentioned 22) that could be identified with 7 of the 11 known migraine-magnesium connections. Thus by choosing a cutoff ratio of 8, the list of 160 search terms could be cut almost in half with a loss of only one of the known connections. Further increases in the cutoff reflected diminishing returns in that respect.

There was considerable overlap between certain sets, so an attempt was made to organize the 84 terms into groups that would yield disjoint or non-overlapping sets. Fourteen such groups were identified and a second Medline search was conducted. The search yielded altogether about 120,000 records in the 84 search sets shown in Table 2. Set numbers corresponding to the fourteen groups are marked with an asterisk; these sets are formed from the union of sets of the individual members of each group.

TABLE 2: DIALOG MEDLINE Search; June, 1989.

Set	Items	Description
S1	467	WEATHER/TI
S2	1362	METOCLOPRAMIDE/TI
S3	2528	ACUPUNCTURE/TI
S4	3908	ASPIRIN/TI
S5	4419	RELAXATION/TI
S6	36	TOLFENAMIC(W)ACID/TI
S7	172	HEPATO(W)BILIARY/TI
S8	179	TEMPERATURE?/TI AND (FINGER OR FOREHEAD OR PERIPHERAL)/TI
S9	78	BLOOD(W)DISORDER?/TI
S10	3114	MENSTRUAL/TI
S11	1072	MOOD/TI OR MOODS/TI
S12	952	((CEREBROSPINAL OR CEREBRO(W)SPINAL)(W) FLUID OR CSF)/TI AND (ACID OR ACIDIC OR BASE OR BASIC)/TI
S13*	18215	S1-S12/OR
S14	73	PLATELET(W)DISORDER?/TI
S15	8005	PLATELET?(2N)(FUNCTION OR AGGREGA? OR ACTIV?T?)/TI
S16	3708	MONOAMINE(W)OXIDASE/TI OR MAO/TI
S17	1082	DOPAMINE(W)BETA(W)HYDROXYLASE/TI
S18	396	PROSTAGLANDIN(1W)(INHIBITOR? OR ANTAGONIST?)/TI
S19*	12982	S14-S18/OR
S20*	11826	(5(W)HT OR 5(W)HYDROXYTRYPTAMINE OR SEROTONIN)/TI
S21	123	PIRIBEDIL/TI
S22	324	METHYSERGIDE/TI
S23	1058	LEVODOPA/TI
S24	63	(BC(W)105 OR SANDOMIGRAN)/TI
S25	157	(ANTISEROTONIN? OR ANTI(W)SEROTONIN?)/TI
S26*	1709	S21-S25/OR
S27	287	FLUNARIZINE/TI
S28	3026	CALCIUM(1W)(BLOCK? OR ANTAGONIST?)/TI
S29*	3287	S27ORS28
S30	2623	CLONIDINE/TI
S31	489	BASILAR(W)ARTERY/TI
S32	259	TEMPORAL(W)(ARTERY OR ARTERIES)/TI
S33	371	RHEOENCEPHALOGR?/TI
S34	605	HEMODYNAMIC?/TI AND (BRAIN OR CEREBR? OR CRANIAL OR CEPHAL?)/TI
S35	1268	MITRAL(W)VALVE(W)PROLAPSE/TI
S36	1400	ANGIOGRA?/TI AND (CEREBRAL OR VERTEBRAL)/TI
S37	1142	EXTRACRANIAL/TI
S38	1090	INTRACRANIAL/TI AND (VESSEL? OR VASCULAR OR ARTER?)/TI
S39	4741	STROKE/TI OR STROKE/TI
S40	138	HYPOPERFUSION/TI OR OLIGEMIA/TI
S41	4343	(NEUROGENIC OR NEUROPATH? OR FOCAL OR TRANSIENT OR TRANSITORY OR CEREBR?/TI OR BRAIN)/TI AND (ISCHEMI? ? OR ISCHAEMI? ?)/TI
S42	4299	(PERFUSION OR BLOOD(3N)(FLOW OR CIRCULATION))/TI AND (CEREBR? OR CAROTID OR CRANIAL OR CEPHAL?)/TI
S43*	21669	S30-S42/OR
S44	1049	DIPYRIDAMOLE/TI
S45	1080	RAYNAUD/TI

```
S46   907  VASOSPAS?/TI
S47   243  VASOACTIVE(1W)SUBSTANCE?/TI
S48   317  CARDIOVASCULAR(W)REFLEX?? ?
S49   608  REACTIVITY(3N)(VESSEL? OR VASCULAR OR
           CEREBROVASCULAR)/TI
S51*  4180 S44-S49/OR
S52   120  CHOCOLATE/TI
S53   453  PHENETHYLAMINE/TI OR PHENYLETHYLAMINE/TI
S54   704  TYRAMINE/TI
S55   822  FOOD? ?(5N)(ALLERG? OR SENSITIVITY OR
           HYPERSENSITIVITY)/TI
S56*  2070 S52-S55/OR
S57   952  VERTIGO/TI
S58   984  CONFUSION?/TI
S59   1079 CONSCIOUSNESS/TI
S60   2877 PAROXYSM?/TI
S61   13156 EPILEP?/TI
S62   186  TRANSIENT(W)GLOBAL(W)AMNESIA/TI
S63*  19004 S57-S62/OR
S64   430  (DIHYDROERGOTAMINE OR DHE(W)45)/TI
S65   899  (ATENOLOL OR TENORMIN)/TI
S66   1171 ERGOT?/TI
S67*  2469 S64-S66/OR
S68   6097 INDERAL/TI OR PROPRANOLOL/TI
S69   6951 BETA(2W)BLOCK?/TI
S70*  12660 S68ORS69
S71   10   BIOFEEDBACK/TI AND AUTOGENIC/TI
S72   47   BIOFEEDBACK/TI AND TEMPERATURE/TI
S73   34   BIOFEEDBACK/TI AND (BLOOD OR VASO? OR
           VASC?)/TI
S74*  89   S71-S73/OR
S75   227  ELECTRONYSTAGMOGRAPHIC/TI
S76   814  OPHTHALMOPLEG?/TI
S77   483  (NEURALGIA OR NEUROPATH?)/TI AND (OCULAR
           OR OPTIC)/TI
S78*  1517 S75-S77/OR
S79   1253 EVOKED(W)POTENTIAL/TI
S80   359  SPREADING(1W)DEPRESSION/TI
S81   10571 (EEG OR ELECTROENCEPHALOGRAPH?)/TI
S82   1081 VASOMOTOR/TI OR VASOMOTORIC/TI
S83   547  LATERALITY/TI OR LEFT(W)HANDEDNESS/TI
S84   393  AUTONOMIC(4N)(DYSFUNCTION? OR DISORDER?
           OR IMBALANCE? ?)/TI
S85*  14133 S79-S84/OR
S86   4413 DEFICIENCY/DE AND (INDUCED OR INTAKE OR
           DIET OR DIETARY OR DIETS OR
           SUPPLEMENT?)/TI
S87   37485 DEFICIENCY/DE AND (PD/DE OR ME/DE OR
           DEFICI?/TI OR METABOL?/TI)
S88   34025 87NOT86
S89   4   13AND86
S91   22  13AND88
S92   11  19AND86
S93       19AND88
S94   9   20AND86
S95   0   26AND86
S96   4   26AND88
S97   1   29AND86
S98   6   29AND88
S99   1   43AND86
S100  32  43AND88
S101  2   51AND86
S102  10  51AND88
S103  1   56AND86
S104  7   56AND88
S105  9   63AND86
S106  62  99-105/OR
S107  0   67AND86
S108  1   67AND88
S109  2   70AND86
S110  8   70AND88
S111  0   74AND86
S112  0   74AND88
S113  3   78AND86
S114  16  78AND88
S115  2   85AND86
S116  13  85AND88
S117  45  109-116/OR
```

The union of sets 1-85 contains about 120,000 records. What we want next is a frequency distribution of substantive (causally relevant) words and phrases within these titles. We do not yet know that we are looking for magnesium. However it is instructive to step outside of the procedure and note at this point that in fact there are only 127 occurrences of "magnesium" among these 120,000 — only 1 per 1000, which is no better than random. This does not mean that the approach is necessarily futile, for we may have still have singled out the relevant physiological connections; moreover, individual search sets that make up the composite may contain significantly large numbers of records with "magnesium".

In any event, we know that 120,000 is impractically large so we are motivated to narrow it down by any reasonable means.

Target Search Strategies

The first literature we identified in this exploratory process was the migraine literature itself. The problem of finding a second literature (called here a "target" literature) that is logically related to the first, but noninteractive with it, is virtually open-ended; the target of this search is initially unknown, but is presumed here to reside within the 120,000 intermediate-stage records that resulted from searching title-correlates of "migraine". It is possible of course to guess at a target, then test that guess by forming the intersection with the 120,000 intermediate records. The very large number of possible targets makes such a direct attack impractical. However, it may not be so difficult to test certain entire categories of targets.

For example, one might assume at the outset that the unknown cause of the given disease, such as migraine, is to be found in some exogenous factor introduced, intentionally or accidentally, into the body. The categories of substances clearly of greatest interest would include dietary factors (or their absence — deficiency factors), toxic factors, and drugs — prescribed or otherwise. Thus it would seem reasonable to postulate and test a variety of what we may call "target search strategies". Medline subheadings appear to be especially useful in designing such strategies, for reasons discussed in [4] (p. 36, "Implications for Indexing"), though descriptors and text words can also have an important role. Table 3 shows four strategies that have been developed — based on deficiency states, dietary factors, poisons, and toxicity studies, respectively. Certain other strategies, such as those based on drugs or genetic factors, for example, could readily be developed. (It is not clear, however, that all types of causes or cures necessarily can be subsumed under some general search strategy.)

TABLE 3: Target-search strategies

Abbreviations:
DIETI1 = (intake or diet or dietary or diets or supplement?)/ti
DIETI2 = (feed or fed or feeding)/ti
INHALATI = (inhalation or inhaled or fume or fumes
or vapor or vapors)/ti
df = deficiency/de po = poison?/de;
xx = xx/de where "xx" designates any Medline 2-letter subheading code, and "de" designates "descriptor".
All other words below are to be given the qualifier "/ti".

The number of postings is shown for each strategy. The various levels are given in order of decreasing effectiveness (measured by ratio R/Z for the best 10 of 26 test substances.) (see text). R/Z is shown as a 1 or 2 digit number immediately following the A,B,C.

I. Deficiency factors. 3 levels
A: 35 4400 df and (induced or DIETI1)
B: 23 34000 df and (pd or me or defici? or metabol?) not A
C: 10 10300 df and (bl or tu or pc or et)(not A not B)

II. Dietary factors 2 levels
A: 19 27400 DIETI1 and (pd or tu or pc or bl or me)
B: 6 22400 (DIETI1 or (DIETI2 and (pd or tu or pc or bl or me))) not A

III. Poisoning 3 levels
A: 32 4800 po and (ingest? or exposed or exposure? or induced or DIETI1)
B: 16 13400 po and (ci or pd) not A
C: 13 62000 (po or exposed or exposure or contamina?) (not A not B)

IV. Toxicity 3 levels
A: 25 12000 to and (induced or ingest? or exposure? or exposed or contamina? or INHALATI or DIETI1)
B: 13 51000 (to and (bl or toxicity) or (exposed or exposure? or INHALATI or contamina?)) not A
C: 9 12000 (to and ci) not A not B

The search strategies shown in Table 3 were developed and tested with the aid of a list of 52 specific substances. Half of that list (corresponding to 16 dietary or deficiency factors and 10 toxic substances) was used as a basis for developing the strategies, and the other half of the list was used to test these strategies. The effectiveness of a search was measured in terms of the previously mentioned frequency ratio for ranking terms, a ratio that also can be interpreted as R/Z, where R is "recall" (defining total relevant as the result of a Medline search on the substance name) and Z is the recall that would result from a random selection of the same total number of hits. (One can show that R/Z is also the same as a similar ratio with precision replacing recall.) Using this measure, the search strategies shown in Table II were found to be highly effective when tested against the 26 substances from the second half of the original list — that is, not including any of the 26 that were used to develop the strategies. The R/Z measure shown is the average of the best 10 among the 26 substances searched.

Strategies IA and IB (deficiency factors) in Table 3 were then introduced into the search as sets S86-S88 (Table 2). An intersection was formed with each of the 14 disjoint groups mentioned earlier (sets S13, S19, S20, S26, S29, S43, S51, S56, S63, S67, S70, S74, S78, S85 of Table 1). The intersection of each such group with the set for each strategy (IA, IB) was formed. For ten of the thirteen groups, the intersection for both sets (IA,IB) was used; for sets S19, S20, and S63, only IA was used. The purpose in using the more restrictive strategy for these larger groups was to limit the total number of titles to a manageable number and at the same time to prevent excessive representation of just a few groups in the exploratory process.

The 160 titles for each of the resulting 14 groups (S93-S98, S106, S117) were then downloaded and edited. In the process of editing, only those terms were selected that corresponded to exogenous substances that might represent deficiency states, or drugs or dietary factors used to treat deficiency states; 222 terms were selected. These terms were sorted alphabetically in order to determine the frequency of occurrence of each term. Of greatest interest are the more frequent terms, assuming that the co-occurrence approach being tested has some validity. In order of decreasing number of occurrences within the 160 titles, and with ratio R/Z (frequency within the 160 titles divided by frequency within the Medline database) shown in parentheses, the results are as follows:

13 magnesium (72)	5 vitamin B12 (90)
10 growth hormone (39)	5 protein (2)
8 aspirin (75)	4 potassium (11)
8 clonidine (110)	4 propranolol (25)
7 cytochrome (24)	4 sodium (5)
6 folate OR folic (61)	4 vitamin E (53)
5 thiamine (90)	<4 all the rest

The list is placed in sequence by absolute number of occurrences rather than frequency ratio because the number of different connections associated with each term is more important than the net statistical improbability of the term itself (measured by R/Z), particularly since all but two terms (protein and sodium) have a very high value of R/Z.

Our interest is further focused only on noninteractive literatures. A Medline search of each term in the list indicated that the sets for aspirin, clonidine, and propranolol all intersected the migraine set substantially, (as might reasonably be expected) — thus one can infer that any effect that these three substances have on migraine headache is well known and so of no interest in this study — the point being to find substances with unreported effects. These 3 substances thus would be eliminated from consideration and the remaining substances retained as candidates for further study.

The fact that magnesium occurs at the top of the list appears to be a striking confirmation that the procedure followed is useful in the process of searching for complementary but disjoint literatures. Further support is given by the particular literatures in the search that led to the 13 magnesium titles — namely sets S20, S27, S28, S35, S40, S46, S61, S80, S82. Among these nine sets, six can be identified with a connection shown in Table 1, as follows: S20 with #7, S27 and S28 with #3, S46 with #2, S61 with #5, and S80 with #4. The arguments presented in the medical review [2] showed why these connections are physiologically plausible.

The same process was applied to the Raynaud's disease/fish-oil example [1], but with Strategy II (Table 3) substituted for strategy I. The intersection of these sets with

Strategy IIA led to 215 titles that were then downloaded and indexed for exogenous substances. Again, as in the migraine search, the technique appears to be remarkably successful; the top-ranked term was in fact the treatment factor (fish oils) originally identified and reported [1].

The results of analyzing the third example, on arginine and somatomedins [3], turned out quite differently. Although a form of the "deficiency" strategy resulted in a frequency ratio R/Z = 7.8 for "arginine" (the target factor discovered in [3]), there were only two occurrences in 387 titles. The two occurrences corresponded to just one intermediate, but correct, connection ("growth hormone"). Technically the outcome could be considered a success. Whether it actually would have proved valuable as an aid to the search is doubtful. An unsuccessful effort was made to find better target search strategies. The nature of the problem can in part be appreciated by examining the list of references in [3]. These titles do not reveal the logical connections of interest to the same degree as in the migraine and Raynaud's disease examples.

Limitations of the Test, and a Proposed Improvement

It should be understood that a retrospective test of the above kind has value principally in suggesting improved approaches to new problems; thus it should be interpreted as an exploration and not as an attempt to validate the proposed method. The method described above was applied only after the "answer" had been developed from the earlier analyses [1, 2]. The possibility of unintended bias in the title- editing process and in the construction of search terms cannot be excluded.

The selection of Strategy I for the migraine example and Strategy II for the Raynaud's disease example requires comment, for these choices were informed by advance knowledge of the outcome. For any new example, one would have to try each of the four strategies one at a time, thus implying that extra work would be necessary in evaluating the other three, presumably unsuccessful, strategies.

Finally the fact that the method appeared to be of little use in the case of [3] suggests that there may be certain classes of problems that are not amenable to such an approach.

The problems encountered with the foregoing method suggest an alternative approach to the search problem that may be more promising. Instead of indexing titles and generating the large intermediate searches such as shown in Table 1, one can take more direct advantage of the target-search strategies such as given in Table 3. Two sets of titles would be downloaded — the original disease-word titles such as "migraine", and the target search strategy titles. A fully automatic procedure (based on a stoplist) can then discover all physiologically-significant words and phrases common to these two sets. I have tested with some success a "stoplist" of about 2000 words (thus far) that can be excluded as candidates for interesting connections. It became apparent in compiling this list that certain words are of no interest if considered singly, but could be of great interest if their immediate context is taken into account. Thus a "go-list" of word pairs or triplets that appear promising can also be used for text selection and should take precedence over the stoplist.

For each word or phrase that the two sets have in common, taken in succession one at at time, the corresponding titles or abstracts in the two sets can then be displayed together in what amounts to a flexible automated browsing system. Prior to that process, the common words and phrases can be grouped either by frequency or, perhaps preferably, by similarity of meaning. This new procedure would make it easier to identify logically-related titles because candidate records from the two literatures would be brought into proximity. The point of such a procedure is to use the computer to manipulate and display titles, abstracts, and/or fragments of text in a way that stimulates human ability to see new connections and to form new hypotheses.

Conclusions

1. Identifying CBD literatures, taking an authentic medical or scientific problem as a point of departure, requires an interactive process in which human judgment must be applied at several points. There is some exploitable information in word frequency distributions, as judged from a study of titles, but not enough to encourage the attempt to develop a fully automatic process.

2. Various target-search strategies were developed and tested; these depended strongly on the use of Medline subheadings. Subheadings are clearly valuable, even though at present limited in scope and not consistently applied. Improvement and strengthening of this aspect of NLM indexing would be of great value to the process described here of exploratory searching in a quest for clues to causal connections. Although the use of target-search categories introduced certain limitations with respect to objectives, the probability of success within those limits was greatly increased. Such categories or strategies, as used in this analysis, represented in effect hypotheses about categories of causal factors, or possible treatments, for the disease under investigation. There is probably no substitute for starting with good hypotheses.

3. A lexical approach in which a stoplist of non-causal words is identified in advance appears promising as a component of an automated browsing system.

Related Work

The work reported here is part of an ongoing series of studies, whose kinship and ancestry have always been something of a puzzle. Work on interfield theories seems closely related [7]; numerous studies in bibliometrics and

other aspects of information science are cited in [4]. Roy Davies has traced the ancestry of work similar to this to roots in the 19th century, pointing out numerous connections of which I had been unaware; he also suggests additional structures and strategies to supplement those that I have described [17].

Acknowledgement and Comment

I am grateful to Abe Bookstein and David Lewis for valuable suggestions and helpful discussions.

I am also appreciative of comments by two anonymous referees who questioned whether this work has much to do with IR. This concern prompted much revision, and, I hope, improvement. In one important respect, however, the same question remains. This work is certainly outside of the usual "user" framework. There is no identifiable community of "users", and no sets of "test questions" have been either designed or collected. Moreover, there seems to be no need for "relevance judges" to do whatever such people are supposed to do, a point I admit to never having clearly understood anyhow.

References

1. Swanson, Don R. "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge". *Perspectives in Biology and Medicine*. 30(1):7-18; 1986.

2. -------- "Migraine and Magnesium: Eleven Neglected Connections." *Perspectives in Biology and Medicine*. 31(4):526-557; 1988.

3. ------- "Somatomedin C and Arginine; Implicit Connections Between Mutually-Isolated Literatures". *Perspectives in Biology and Medicine*. 33(2):157-186; Winter, 1990.

4. ------- "Medical Literature as a Potential Source of New Knowledge". *Bulletin of the Medical Library Association* 78(1):29-37; 1990.

5. ------- "The Absence of Co-Citation as a Clue to Undiscovered Causal Connections." In: C. L. Borgman, ed. *Scholarly Communication and Bibliometrics*; Newbury Park, CA; Sage Publications, 1990; p. 129-137

6. DiGiacomo, R. A. et. al. "Fish-Oil Dietary Supplementation in Patients with Raynaud's Phenomenon: A Double-Blind, Controlled, Prospective Study. *American Journal of Medicine* 86:158-164; 1989.

7. Ramadan, N. M., et. al. Low Brain Magnesium in Migraine. *Headache* 29:416-419; 1989.

8. Taubert, Konrad and Gerd Keil. "Pilotstudie zur Magnesiumtherapie bei Migræne und Spannungskopfschmerz" *Zeitschrift fuer Ærztliche Fortbildung*. 85:67-8; 1991.

9. Sutton, Walter S. " The Chromosomes in Heredity". *Biological Bulletin* 4 n5:231-251; 1903.

10. Darden, Lindley. and Maull, Nancy. "Interfield Theories". *Philosophy of Science* 44:43-64; 1977.

11. Mendel, Gregor. "Experiments on Plant Hybrids" [1865] in C. Stern and E. R. Sherwood, eds. *The Origin of Genetics*. San Francisco: Freeman, 1966; p. 1-48.

12. Swanson, Don R. "A Second Example of Mutually-Isolated Medical Literatures Related by Implicit, Unnoticed Connections." *Journal of the American Society for Information Science*. 40(6):432-435; 1989.

13. ------ "Online Search for Logically-Related Noninteractive Medical Literatures: A Systematic Trial-and-Error Strategy." *Journal of the American Society for Information Science*. 40(5):356-358; 1989.

14. Wimsatt, William C. "Robustness, Reliability, and Overdetermination" in M. B. Brewer & B. E. Collins, *Scientific Inquiry and the Social Sciences*. San Francisco: Jossey-Bass; 1981. p. 124-163.

15. ------- "Heuristics and the Study of Human Behavior" In: D. W. Fiske & R. Schweder, eds. *Metatheory in Social Science. Pluralisms and Subjectivities*. Chicago: U. Chicago Press; 1986. p. 293-314.

16. Swanson, Don R. *Medical Literature as a Source of New Knowledge: USDE Grant R039A80028 FINAL REPORT*. Chicago: University of Chicago Grad Lib Sch and Center for Information and Language Studies. Dec. 1989.

17. Davies, Roy. "The Creation of New Knowledge by Information Retrieval and Classification" *The Journal of Documentation* 45(4):273-301; 1989.

An earlier phase of this work received support from the USDE Office of Educational Research and Improvement. Grant R039A80028; 1988/1989.