Pergamon

Technovation 21 (2001) 689–693

technovation

www.elsevier.com/locate/technovation

Predicting emerging technologies with the aid of text-based data mining: the micro approach

N.R. Smalheiser *

Department of Psychiatry, University of Illinois at Chicago, MC912, 1601 W. Taylor Street, Chicago, IL 60612, USA

Accepted 11 April 2001

Abstract

Text data mining should be useful for anticipating new technologies and new uses for existing technologies, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature. The present study attempted to predict genetic engineering technologies that may impact on viral warfare in the future. The analysis was carried out using a combination of conventional Medline searches and the package of advanced informatics techniques known collectively as Arrowsmith. The findings strongly indicate that genetic packaging technologies such as DEAE-dextran, cationic liposomes and cyclodextrins are plausible candidates to enhance infections caused by viruses delivered via an aerosol route — despite the fact that no studies have yet been reported that have examined this issue directly, and certainly not in the contexts of viral disease or viral warfare. The critical factor was the overall strategy of approaching the problem: first, to define two specific fields explicitly (in this case, genetic engineering and viral warfare) that are hypothesized to contain complementary information; second, to identify common factors that bridge the two disciplines (i.e. research on viruses); and third, to progressively shape the query once initial findings are obtained. Thus, in contrast to some current perceptions, the process of text data mining is neither automatic nor is it restricted to those who have access to macro analyses using customized computer systems.

Keywords: Informatics; Information retrieval; Arrowsmith; Forecasting; Prediction

1. Introduction

Technological innovation often proceeds by applying advances made in one field to a separate arena. Once the innovation is implemented, the transfer of knowledge may appear obvious or even inevitable, but without the benefit of hindsight it is surprisingly difficult to identify specific technologies that are ripe for transfer. One must simultaneously identify a need in one domain and a tool in another, possibly quite disparate domain that potentially satisfies that need — and such a task requires more than expert knowledge. A large body of research knowledge is published in the form of papers and technical reports that are accessible via bibliographic databases, leading several workers to advocate the development of techniques for knowledge discovery in databases (Fayyad and Uthurusamy, 1999), and in particular, strategies for text data mining (Swanson and Smalheiser, 1997; Hearst, 1999; Kostoff, 1999), in order to 'discover' useful knowledge that is implicit within the published record. Anticipating new technologies and new uses for existing technologies should be ideal applications for text data mining, insofar as one can attempt to connect complementary pieces of information across two different domains, or subsets, of the scientific literature, that may not have been noticed by workers beforehand.

Text data mining strategies can be divided into two types, macro and micro. Macro analyses perform data-crunching operations over a large, often global set of papers encompassing one or more fields, in order to identify large-scale trends or to classify and organize the literature. Several examples of macro analyses have been published by ourselves and others (Swanson and Smalheiser, 1997; Kostoff, 1999). In contrast, micro analyses pose a sharply focused question, in which one

* Tel.: +1-312-413-4581; fax: +1-312-413-4569.
E-mail address: smalheiser@psych.uic.edu (N.R. Smalheiser).

0166-4972/01/$ - see front matter
PII: S0166-4972(01)00048-7

searches for complementary information that links two small, pre-specified fields of inquiry. We have previously shown the value of this micro approach in helping to formulate and assess hypotheses arising in biomedical research (Smalheiser and Swanson, 1998a,b), and in the present paper, it is demonstrated how the micro approach can be employed for helping make policy decisions regarding technical innovation.

Genetic engineering technologies have the capability to alter the make-up of biological organisms and thus have the potential to impact on the way that nations may conduct, and hopefully may defend against, the threat of biological warfare (BW). To anticipate possible threats that may be developing, one needs to learn what relevant genetic research is being done around the world — not only research that is explicitly intended for military applications, but also research being conducted in medical, biotechnological, public health, agricultural or zoological contexts that might be potentially applied to BW applications in the future. This is a task for military intelligence, but it can be difficult to distinguish research intended for BW from that directed toward, for example, vaccine development or gene therapy, and intelligence officers need to prioritize which kinds of genetic research are most in need of being tracked. The task is made even more difficult by the multiplicity of BW scenarios that must be considered — for example, whereas battlefield deployment of BW agents would necessarily induce acute, fulminant disease that incapacitates troops, a terrorist threat might well involve dissemination of agents that induce chronic rather than acute symptoms.

Taking the complementary approach of using informatics to predict emerging genetic engineering technologies that may impact on BW, specifically viral warfare, the question is "Given the state of published research right now, what BW applications are possible?" whether or not there is evidence that anyone is actually exploring those avenues. The analysis was carried out using a combination of conventional Medline searches and the package of advanced informatics techniques known collectively as Arrowsmith (Swanson and Smalheiser, 1997), which seek to find meaningful relationships between two largely disparate literatures or fields of inquiry — in this case, genetic engineering vs. viral warfare. Furthermore, the focus was on research findings that were so strong and consistent that they were reflected directly in the titles of papers, although the abstracts and text of key papers were also assessed when relevant.

2. Results

The first step in this analysis was to define the problem more precisely and narrowly, in order to define two subsets of the literatures on viral warfare and genetic engineering that are likely to be implicitly related to each other in a complementary fashion. Because aerosol dispersion of viruses is a major scenario for viral warfare, and because viruses must remain stable in aerosols for a significant period of time if they are to be regarded as warfare threats, it was decided to focus initially on studies of the aerosol stability of viruses. Furthermore, the entire field of genetic engineering was not examined, but initial attention was restricted to studies that had examined virulence of viruses using genetic techniques. Thus, this approach was intended to identify ways of making viruses already known to be virulent even more effective in aerosol attacks, by altering their genetic make-up so as to increase their aerosol stability. However, the information may apply to other scenarios as well, for example, making non-virulent viruses more virulent; using viruses as vectors to carry exogenous genes that encode toxins; or dispersing viral nucleic acids, rather than intact viruses, by an aerosol route as a way of infecting hosts.

The second step in the analysis was to retrieve the existing literature on ways to alter the aerosol stability of viruses. A conventional Medline search using the public PubMed search engine (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=PubMed; August 2000) posed the query "(aerosol or aerosols) AND (stability or viability or survival) AND (virus or viral)", resulting in ~200 records that were inspected manually. As an alternative method of identifying relevant papers on this topic, an Arrowsmith search was carried out where one literature A=papers dealing with aerosol stability/survival of viruses and a second literature C=genetics/genetic techniques and virulence of viruses; title words that were shared in both literatures were denoted as B-terms and we examined the records whose titles contained B-terms that corresponded to the names of individual viruses (see details in Swanson and Smalheiser, 1997). Both approaches gave similar results: Previous research has shown that the aerosol stability of viruses can be altered by varying the relative humidity and the temperature of the air, as well as by adding compounds such as DMSO, inositol and polyethylene glycol which retain adherent water and prevent dehydration of the virus with subsequent denaturation of viral proteins (a critical factor in loss of virulence in aerosols). No indication was evident from the current literature that manipulating the genetic make-up of intact viruses could affect their aerosol stability. However, a series of papers was found showing that the aerosol stability of naked nucleic acids including viral DNA and RNA can be enhanced using packaging technologies such as the use of DEAE-dextran or cationic liposomes, that are used extensively for gene transfer experiments.

This led us to revise the query, to assess whether one or more of these packaging technologies might plausibly

affect intact viruses during aerosol attack. The scope was broadened, because the overall outcome of the effects of viral infection is the most relevant parameter for BW, and includes not only aerosol stability but also initial infection of the virus for airway cells, gene expression, replication, and subsequent spread and adverse effects upon the host. Three popular, yet quite different packaging technologies for examination were chosen: DEAE-dextran, cationic liposomes, and cyclodextrins, and PubMed searches were carried out on each to learn what research had been done already in terms of aerosol delivery (of anything — drugs, proteins, nucleic acids or viruses) as well as in the context of viruses.

2.1. DEAE-dextran

A PubMed search of 'DEAE-dextran and virus' yielded 263 records, of which 29 were inspected as possibly relevant based on titles. These indicated that many studies have found that, in many cases though not universally, DEAE-dextran significantly improves the infectivity and net gene expression of intact viruses (not just naked nucleic acids) in cultured mammalian cells in vitro — and in several cases, were tested and found effective in vivo as well. Two papers were especially noteworthy, because they showed that expression of viral transgene expression was enhanced by DEAE-dextran in transfection of airway epithelial cells both in vitro (Arcasoy et al., 1997) and instilled intranasally in vivo (Kaplan et al., 1998). In contrast, however, a search on 'DEAE-dextran and aerosol' yielded only two papers, neither particularly relevant to the query except that they demonstrate the point that DEAE-dextran can be applied to lungs in aerosol form (Barrowcliffe et al., 1990). A search on 'DEAE-dextran and virulence' retrieved 39 records, most of which dealt with infection of cells in vitro or transfection of viral nucleic acids rather than intact viruses, but a few suggested that DEAE-dextran packaging could enhance viral infection in vivo as well. Searches were also carried out on the alternative terms 'polycations and virulence', 'polycations and virus', 'polycations and aerosol', and 'polycations and airway', in order to detect additional papers that might lack the specific term DEAE-dextran in the Medline records, and the results were consistent with the analysis above.

To summarize, there is an abundance of evidence that DEAE-dextran and other polycations can enhance gene transfer of intact viruses in vitro, and one recent key paper (Kaplan et al., 1998) suggests that DEAE-dextran can enhance virally derived gene expression when viruses are instilled intranasally in vivo. These findings suggest that it is plausible to hypothesize that DEAE-dextran may facilitate virus infection induced by an aerosol route by acting at one or more steps. Note that this suggestion is novel, since no studies were found that examined effects on net infectivity or virulence of intact viruses applied in aerosol form, and only a few studies (more than 20 years old) even examined the effects of DEAE-dextran on viral infection in vivo. However, in the text of the Kaplan et al. (1998) paper, it is stated that "Studies are currently underway to generate an aerosol formulation of Adenovirus/polycation complex for delivery to the lung." Evidently research is underway, albeit not yet published, on the premise that DEAE-dextran is a promising agent for aerosol delivery in the context of gene therapy/gene transfer.

There are at least three methodologic limitations on the above searches: (a) Incomplete indexing may be a factor, and one paper was found where DEAE-dextran was mentioned in the text of the paper but not in the Medline record (De Jong et al., 1974) was found. In that paper, DEAE-dextran was found to have mixed effects: it lowered the titer of infectious intact virus somewhat, but also helped to maintain the infectivity of aerosol-denatured viral RNA. (b) Alternative word usage. Additional records were found using the term 'polycations', reflecting the fact that a variety of polycations has been utilized in these experiments besides DEAE-dextran. Although inspecting these records only reinforced the findings found using the term 'DEAE-dextran', conceivably some relevant papers related to other specific polycations were still missed. (c) Changing importance of DEAE-dextran within the scientific community over time. Whereas DEAE-dextran was one of the first packaging agents known to aid transfection in gene transfer, more recently cationic liposomes have been popular because they are in some cases more effective and are applicable for additional applications (e.g. to package proteins and drugs as well as viruses and nucleic acids). Thus, more recent papers might be expected to give more emphasis to cationic liposomes; but in fact, as discussed below, similar findings were obtained when considering the latter packaging technology as well.

2.2. Cationic liposomes

A PubMed search of 'cationic liposome(s) and virus' retrieved 239 records, of which 30 titles were inspected as possibly relevant based on titles. Cationic liposomes have been investigated in a wide variety of formulations and protocols for gene transfer (Colosimo et al., 2000). In particular, cationic liposomes enhance infection of a variety of cell types (including airway epithelium) by a variety of viruses (adenoviruses and retroviruses have been best studied since they are vectors for gene therapy). In some cases, the enhancement in infection appears to be due to enhanced uptake of the virus by cells, either via its normal route or bypassing its normal cellular receptors (Innes et al., 1990; Bass et al., 1992; Qiu et al., 1998), but enhanced release of virus from endocytic vesicles may also be a factor. As well, cationic

liposomes may aid infection by shielding virus from neutralizing antibodies (Chillon et al., 1998). There are also many examples in which cationic liposomes have been used for aerosol delivery of nucleic acids, either naked or packaged together with defective viruses; these viruses were used as carriers for co-transfection of exogenous naked DNA (Yonemitsu et al., 1999). Nevertheless, we could not find any examples in which cationic liposomes were examined for their effects on aerosol stability of intact viruses, or infection of virus per se that was given by an aerosol route, or virulence of viruses known to cause disease via aerosol transmission.

Thus, the current literature strongly supports the hypothesis that cationic liposomes potentially might have an impact on viral warfare, by affecting one or more steps in viral transmission in aerosols. Cationic liposomes might improve the aerosol stability of certain viruses, by delaying the dehydration of viral coat proteins and/or by preserving the infectivity of viral nucleic acid. As well, cationic liposomes might shield viruses from neutralizing antibodies and facilitate the uptake of viruses by airway epithelium cells. It is also possible that cationic liposomes would allow viruses to bypass host range limitations due to specificity of viral receptors — for example, a virus that normally infects only animal species may be induced initially to infect human airway cells. Although such viruses would not be able to secondarily infect human cells beyond the airway epithelium, it should be noted that even non-replicating viruses can produce pneumonia when introduced into the lung (Zsengeller et al., 1997). Moreover, non-replicating viruses could be engineered to produce toxins within airway cells, and the toxins could spread throughout the body via the bloodstream, even if the viruses themselves could not.

In the face of all this supporting evidence, the question is why have cationic liposomes not been investigated in the context of virus infections acting via an aerosol route. The most likely answer, based on the indication from DEAE-dextran cited above, is that such research is ongoing at present in the context of viruses employed as vectors for gene therapy, albeit the relevant papers have simply not been published yet because the area of scientific activity is so new. On the other hand, one would not expect medical virologists to investigate such artificial means of enhancing viral infections, since they are primarily interested in the natural course of viral diseases (except, perhaps, in the context of occupational hazards of genetics laboratories).

2.3. Cyclodextrins

These have been investigated primarily as carriers for drug delivery, particularly to aid in solubilization and stabilization of drugs in aqueous solutions, including nasal sprays. One report indicates that beta-cyclodextrins are effective in enhancing adenovirus-mediated gene delivery into intestinal epithelial cells in vitro (Croyle et al., 1998), and cationic derivatives of beta-cyclodextrins have also been shown to be effective for transfection of plasmid DNA (Gonzalez et al., 1999). Alpha-cyclodextrin has been reported to preserve the activity of trypsin after spray-drying onto a surface, presumably by preventing its dehydration and subsequent denaturation (Millqvist-Fureby et al., 1999); this raises the possibility that it may also stabilize viruses in aerosols.

3. Discussion

The analysis presented here indicates strongly that genetic packaging technologies such as DEAE-dextran, cationic liposomes and cyclodextrins are plausible candidates to enhance viral infections via an aerosol route, acting by one or more steps — despite the fact that no studies have yet been reported that have examined this issue directly, and certainly not in the contexts of viral disease or viral warfare. Thus, our analysis was not simply an exercise in summarizing the known state of the art, but rather gleaned previously unexamined ‘nuggets’ of potential technology transfer from a mountain of raw data scattered in the current literature (Hearst, 1999) — without the need for large-scale computing, and without relying on expert knowledge of the fields in question.

Although Arrowsmith software was employed to help define and juxtapose the two literatures in question, these advanced programs were not essential, and indeed we found in this case that the same outcome was obtained using conventional Medline searching techniques alone. The critical factor was the overall strategy of approaching the problem: first, to define two specific fields explicitly (in this case, genetic engineering and BW) that are hypothesized to contain complementary information; second, to identify common factors that bridge the two disciplines (i.e. research on viruses); and third, to progressively shape the query once initial findings are obtained. Thus, in contrast to some current perceptions, the process of text data mining is neither automatic nor is it restricted to those who have access to customized computer systems.

Nevertheless, given the broad scope of the field of genetic engineering, it is likely that the present example is but one among many other examples of potential technology transfer from the field of genetic engineering to the field of BW. To gain a more comprehensive and systematic overview of this issue it will be necessary to undertake a macro analysis of the two literatures, e.g. using Arrowsmith to identify a series of items that link the two literatures as a whole, followed by micro analyses to examine these items individually (Swanson and Smalheiser, 1997). Thus, the full power of text data mining is probably best captured by coupling macro analyses with micro analyses.

Acknowledgements

Supported by a grant from the Office of Naval Research to N.R.S., and a contract from the Defense Intelligence Agency to Don R. Swanson.

References

Arcasoy, S.M., Latoche, J.D., Gondor, M., Pitt, B.R., Pilewski, J.M., 1997. Polycations increase the efficiency of adenovirus-mediated gene transfer to epithelial and endothelial cells in vitro. Gene Ther. 4, 32–38.

Barrowcliffe, M.P., Zanelli, G.D., Ellison, D., Jones, J.G., 1990. Clearance of charged and uncharged dextrans from normal and injured lungs. J. Appl. Physiol. 68, 341–347.

Bass, D.M., Baylor, M.R., Chen, C., Mackow, E.M., Bremont, M., Greenberg, H.B., 1992. Liposome-mediated transfection of intact viral particles reveals that plasma membrane penetration determines permissivity of tissue culture cells to rotavirus. J. Clin. Invest. 90, 2313–2320.

Chillon, M., Lee, J.H., Fasbender, A., Welsh, M.J., 1998. Adenovirus complexed with polyethylene glycol and cationic lipid is shielded from neutralizing antibodies in vitro. Gene Ther. 5, 995–1002.

Colosimo, A., Goncz, K.K., Holmes, A.R., Kunzelmann, K., Novelli, G., Malone, R.W., Bennett, M.J., Gruenert, D.C., 2000. Transfer and expression of foreign genes in mammalian cells. Biotechniques 29, 314–331.

Croyle, M.A., Roessler, B.J., Hsu, C.P., Sun, R., Amidon, G.L., 1998. Beta cyclodextrins enhance adenoviral-mediated gene delivery to the intestine. Pharmacol. Res. 15, 1348–1355.

De Jong, J.C., Harmsen, M., Trouwborst, T., Winkler, K.C., 1974. Inactivation of encephalomyocarditis virus in aerosols: fate of virus protein and ribonucleic acid. Appl. Microbiol. 27, 59–65.

Fayyad, U., Uthurusamy, R., 1999. Data mining and knowledge discovery in databases: introduction to the special issue. Commun. ACM 39 (11).

Gonzalez, H., Hwang, S.J., Davis, M.E., 1999. New class of polymers for the delivery of macromolecular therapeutics. Bioconjugate Chem. 10, 1068–1074.

Hearst, M.A., 1999. Untangling text data mining. In: Proc. ACL '99: the 37th Annual Meeting of the Association for Computational Linguistics. Available at http://www.sims.berkeley.edu/~hearst.

Innes, C.L., Smith, P.B., Langenbach, R., Tindall, K.R., Boone, L.R., 1990. Cationic liposomes (Lipofectin) mediate retroviral infection in the absence of specific receptors. J. Virol. 64, 957–961.

Kaplan, J.M., Pennington, S.E., St. George, J.A., Woodworth, L.A., Fasbender, A., Marshall, J., Cheng, S.H., Wadsworth, S.C., Gregory, R.J., Smith, A.E., 1998. Potentiation of gene transfer to the lung by complexes of adenovirus vector and polycations improves therapeutic potential. Hum. Gene Ther. 9, 1469–1479.

Kostoff, R.N., 1999. Science and technology innovation. Technovation 19, 593–604.

Millqvist-Fureby, A., Malmsten, M., Bergenstahl, B., 1999. Spray-drying of trypsin — surface characterization and activity preservation. Int. J. Pharmaceutics 188, 243–253.

Qiu, C., De Young, M.B., Finn, A., Dichek, D.A., 1998. Cationic liposomes enhance adenovirus entry via a pathway independent of the fiber receptor and alpha-(V) integrins. Hum. Gene Ther. 9, 507–520.

Smalheiser, N.R., Swanson, D.R., 1998a. Using Arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. Comp. Meth. Prog. Biomed. 57, 149–153.

Smalheiser, N.R., Swanson, D.R., 1998b. Calcium-independent phospholipase A_2 and schizophrenia. Arch. Gen. Psychiat. 55, 752–753.

Swanson, D.R., Smalheiser, N.R., 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 91, 183–203.

Yonemitsu, Y., Kaneda, Y., Muraishi, A., Yoshizumi, T., Sugimachi, K., Sueishi, K., 1999. HVJ (Sendai virus)-cationic liposomes: a novel and potentially effective liposome-mediated technique for gene transfer to the airway epithelium. Gene Ther. 4, 631–638.

Zsengeller, Z.K., Boivin, G.P., Sawchuk, S.S., Trapnell, B.C., Whitsett, J.A., Hirsch, R., 1997. Anti-T cell receptor antibody prolongs transgene expression and reduces lung inflammation after adenovirus-mediated gene transfer. Hum. Gene Ther. 8, 935–941.

Neil R. Smalheiser is a Research Assistant Professor in the Department of Psychiatry at the University of Illinois at Chicago. His experimental research concerns the role of extracellular matrix proteins in nervous system development, function and disease. He has collaborated with Don Swanson on the Arrowsmith text data mining project for the past seven years.