The Place of Literature-Based Discovery in Contemporary Scientific Practice

Neil R. Smalheiser and Vetle I. Torvik

Abstract In this brief essay, we consider some of the lessons that we learned from our experience working with the Arrowsmith consortium that may have implications for the field of literature-based discovery (LBD) as a whole.

Keywords: Literature-based discovery · Informatics · Text mining · Hypothesis generation

1 Introduction

For the past 5 years, the Arrowsmith consortium has developed a suite of web-based informatics tools to assist biomedical investigators in making discoveries and establishing collaborations. Researchers working in multi-disciplinary neuroscience research groups have served as field testers, and feedback arising from their use of the tools in their daily work has contributed crucially to the project. We have recently described the evolution of the two-node search interface [1], discussed the role of field testers in detail [2], and described a quantitative model for ranking B-terms according to their likely relevance for linking two disparate sets of articles in a meaningful manner [3]. These and other references are available for download on the public UIC Arrowsmith website (http://arrowsmith.psych.uic.edu). Here, we would like to consider some of the lessons that we learned that may have implications for the field of literature-based discovery (LBD) as a whole.

First, what is included in the term literature-based discovery? Most authors who have used the term have referred to the so-called "one-node" or open-ended search, in which a scientific problem is represented by a set of articles (or literature) that discusses the problem, and the goal is to find some other (generally disjoint) set

N.R. Smalheiser and V.I. Torvik
UIC Psychiatric Institute MC912, 1601 W. Taylor Street, Chicago, IL 60612, USA
neils@uic.edu

P. Bruza and M. Weeber (eds.), *Literature-based Discovery,*
Springer Series in Information Science and Knowledge Management 15.

of articles containing information that can contribute to the solution of the problem [4–17]. The Arrowsmith consortium has focused primarily on the "two-node" search, in which a scientist wishes to find or assess links that connect two different sets of articles (again, generally disjoint and in different disciplines) [4, 18]. Don Swanson has proposed the term "undiscovered public knowledge" to refer to the overall process of assembling different bits of knowledge that are scattered across different literatures into a novel hypothesis [19]. Smalheiser has published numerous examples that fall more specifically into the category of "gap analysis" – that is, not so much proposing new solutions to an existing problem, but rather identifying new and potentially important scientific problems that no one seems to be studying or even noticing, either because they fall in the cracks between disciplines or for other sociological reasons [20–24]. Some LBD analyses consider discrete problems, e.g., dietary restriction in aging [25], whereas other analyses comprise more global analyses of entire disciplines, e.g., fullerene research [26]. Some studies make "incremental" predictions such as expanding the list of diseases that can be treated by a given drug [14], whereas some analyses find connections between disparate disciplines (e.g., gene therapy vs. bioterrorism) that have few articles or practitioners in common [23].

Regardless of the particular type or flavor of LBD that is pursued by different individuals, all share a more ambitious agenda than simply to extract or process the information present in a given text. If much of the research in "text mining" seeks to identify relationships that are explicitly stated, then LBD goes further to identify relationships that are implicitly stated – and not within a single document, but across multiple documents. This is a form of "data mining," but most data mining seeks to identify valid relationships within the data, whether or not they have ever been observed previously. In contrast, LBD practitioners have tended to focus on a search for relationships that are entirely novel, never noticed and perhaps never even speculated upon by scientists previously. Thus, the LBD field has set its sights on a very high, perhaps an impossibly high standard: true, novel, un-noticed, non-trivial (and generally cross-disciplinary) scientific discoveries.

2 A Case Study

We recently published a bioinformatics analysis predicting that certain genomic repeat elements within human mRNAs, the so-called MIR/LINE-2 repeats, are likely to serve as targets of the small trans-acting noncoding antisense RNA family known as microRNAs [27]. This raised the question whether other repeat elements may also serve as microRNA targets. Because Alu elements are the most common repeats expressed within mRNAs, we focused on these and found that a family of microRNAs do, indeed, appear to target Alu-containing mRNA transcripts [28]. To look for other types of biological relationship(s) we carried out a two-node search between microRNAs and Alu: The microRNA literature consisted of 970 articles, Alu had 2,945, and the intersection was empty (Fig. 1). A total of 1,428 title words

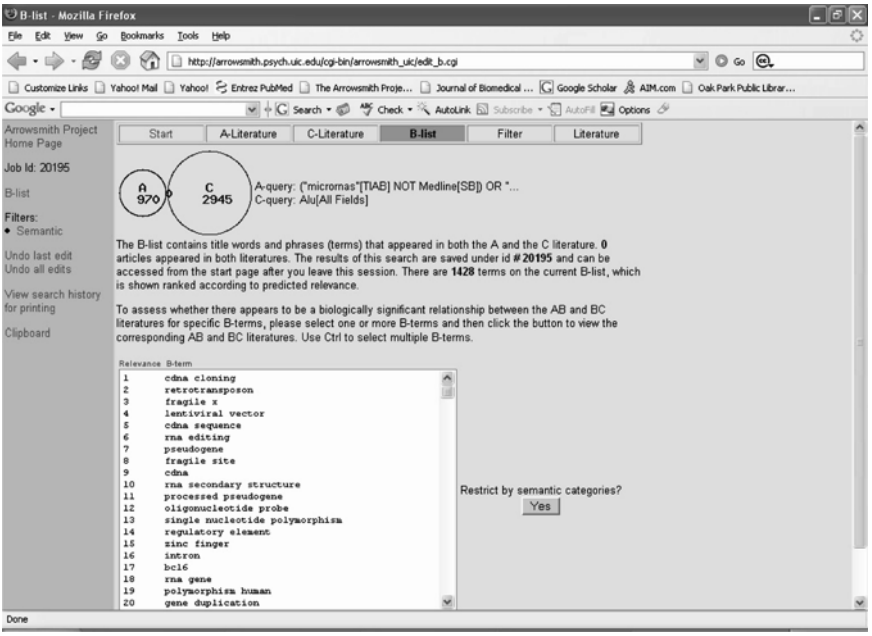

Fig. 1 Screenshot of the two-node search output for the example "microRNA vs. Alu", showing the most highly ranked B-terms

and phrases were in common to the two literatures (B-terms), and these were ranked according to a quantitative model that predicts the terms that are most likely to represent meaningful links across literatures [3]. We examined the top-ranked 100 B-terms, of which the following terms warrant discussion here:

#6, *RNA editing*. Alu repeats are highly edited, particularly within introns of unprocessed mRNAs. As well, some microRNA precursors have been shown to be edited. Finally, extensive RNA editing of a transcript inhibits its ability to be degraded via RNA interference, a pathway of RNA control that overlaps with the microRNA pathway.

#10, *RNA secondary structure*. MicroRNA precursors are characterized by a distinctive hairpin stem-loop structure. Alu repeats also express one or two hairpin loops. This raises the possibility, for example, that they might both bind proteins that recognize hairpin structures.

#25, *differentiation HL-60 cell*. It has been shown that certain microRNAs change their expression during differentiation of HL-60 cells. Separately, a set of transcripts that show significant changes in their subcellular localization and translation during differentiation were found to contain Alu sequences. Could this be a sign that certain microRNAs are targeting Alu sequences within these transcripts?

#33, *RNA binding protein*. MicroRNAs have been reported to associate with FMRP, a RNA binding protein that has an important role in synaptic plasticity. As well, both the Alu-derived small RNA BC200, and the tRNA-derived small RNA

BC1, have been reported to associate with FMRP. On the other hand, cytoplasmic Alu transcripts associate with SRP9/14 and with La/SS-B, which have not been implicated in any microRNA pathways so far.

#54, *antisense RNA*. MicroRNAs are thought to bind to target sequences within the 3′-UTR of mRNAs. A report in the Russian literature suggested that certain riboprotein complexes containing noncoding Alu transcripts may downregulate mRNAs containing Alu elements in the opposite orientation [29] (Fig. 2). If so, this would suggest that inducible Alu transcripts bind certain mRNAs and might be functionally similar to microRNAs (which had not been described in mammalian cells at the time that these papers were published).

The link regarding antisense RNA was particularly intriguing because it pointed to a published series of articles that arguably have gained increasing plausibility and significance in light of the subsequent discovery of RNA interference and of microRNAs. As Don Swanson has said (personal communication), identifying neglected articles worth a second look is a kind of literature-based discovery too! Certainly without carrying out a two-node search we would not have noticed the possibility that cytoplasmic Alu transcripts may bind to Alu-containing mRNAs and regulate their expression via pathways that may be related to RNA interference. This paper [29] and its implications appear to have been neglected despite its indexing in MEDLINE. For example, after our paper was published, Daskalova et al. [30] discussed the possibility that cytoplasmic Alu transcripts may bind to Alu-containing mRNAs, without citing any prior literature on this question.

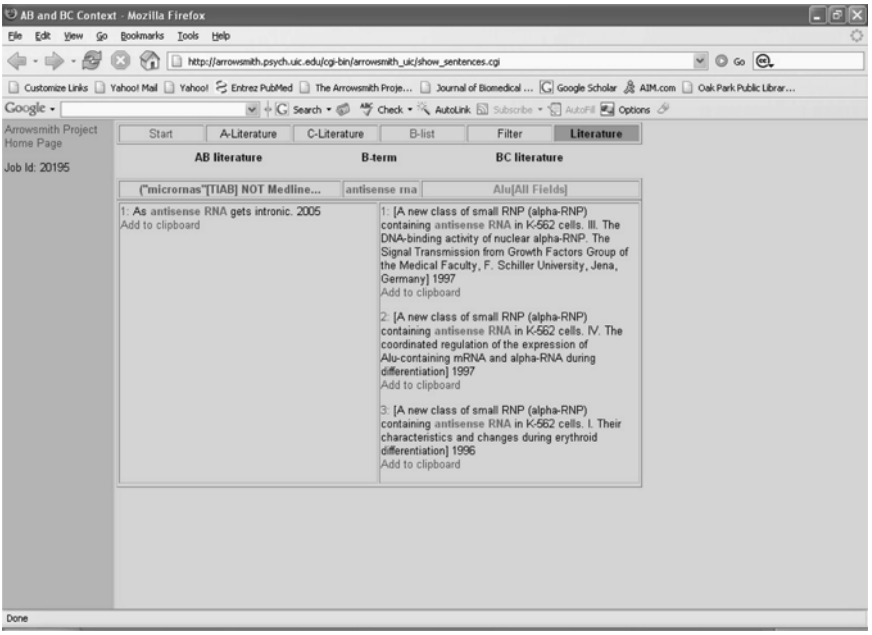

Fig. 2 Screenshot of an AB title juxtaposed to BC titles for the B-term "antisense RNA"

What is the next step? Don Swanson has proposed that a LBD finding can be considered at least partially successful if it leads to publication of a hypothesis paper in the peer-reviewed literature [31]. In the past, we have published individual LBD predictions as short notes in biomedical journals (see references listed in http://arrowsmith.psych.uic.edu). After about a dozen examples had been published, we felt that we had made the point sufficiently that Arrowsmith does assist investigators in generating and assessing hypotheses (another dozen examples of LBD findings were included in a recent description of Arrowsmith field tester behavior [2]). However, publishing papers is not a speedy, and not necessarily an efficient, mechanism for motivating other scientists to test a given hypothesis. For two hypotheses we attempted to alert investigators directly via email, but have either received no reply or short, rather dismissive responses. In principle we could arrange to test the "cytoplasmic Alu transcripts regulating mRNAs" hypothesis ourselves in the laboratory, but these experiments are neither covered by existing grants nor would obviously lead to the writing of any new grant proposals. Despite being plausible, the hypothesis remains an Orphan in search of a Daddy Warbucks.

The moral? One should not assess the value of a LBD prediction according to whether it is tested experimentally, since many pragmatic and (from the LBD viewpoint) irrelevant factors affect whether one is able to carry out an experimental test. As well, LBD is based on an analysis of the structure of the scientific literature, which reflects human activity and does NOT necessarily reflect the structure of nature – so for better or worse, one should not assess the value of LBD predictions according to whether they eventually turn out to be true after all.

3 Re-defining Success in LBD

This case is not necessarily a tragedy: After all, most hypotheses do not survive scientific scrutiny, and most are not even important enough to test at all. Yet the LBD field has defined its own success in terms of whether the hypotheses generated by LBD searches are not only truly novel and significant, but whether they have been tested by others, and whether they were validated experimentally. Given that most people who are involved with LBD do not have laboratory facilities available to test their own hypotheses, this definition of success is almost impossibly high to fulfill, and orders of magnitude beyond what is expected for any traditional search engine or IR strategy.

Let's look again at the case study in terms of what it did right: This two-node search was natural to formulate in the course of ongoing studies of microRNAs and Alu elements. It took seconds to enter the query, less than a minute to return the ranked B-list, and about a half-hour to examine 100 top ranked B-terms for titles and abstracts of the interesting papers. Thus, it did not require a large commitment of time and energy to examine, and could readily be integrated into normal workflow. The search returned non-trivial links between microRNAs and Alu elements. Although we were aware of some of these links already (and were discussed in [28]),

a different person doing this search might have learned new information. Finally, the search readily identified testable and truly inter-disciplinary links between two different disparate literatures, i.e., between the microRNA field and the Alu field.

This suggests that the scope of LBD might be expanded to embrace the full continuum of information that can be retrieved from searches, from retrieval of explicitly stated information, to retrieval of implicit links, to truly novel hypotheses. Similarly, we have found that the actual information-seeking activities of field testers have completely blurred what (to an information scientist) are clear and fundamental distinctions between simple fact-finding, browsing a new literature, and carrying out one and two-node searches [2]. It should not be a surprise that end-users envision cultural products differently than do the developers – millions of people use Google without knowing how the search engine works, and billions of people enjoy music without knowing the rudiments of music theory. If LBD tools are to become popular as well, they need to be usable by people who do not know how they work. It is appropriate for the developers of LBD tools to focus on the procedural aspects and formal methodology of one-node and two-node search strategies, but to the end-users, LBD searches should appear to be simple extensions of simple PubMed searches. And, just as the end-users have blurred the distinctions between different types of information-seeking activities, so might the LBD field benefit from integrating LBD tools with other informatics resources, so that LBD comprises only one part of a larger multi-purpose tool kit. From this standpoint, success is not measured in terms of number of discoveries made, but in how many end-users utilize a given tool and how often.

4 Gold Standards

Within the community of LBD tool developers, perhaps the biggest stumbling-block to progress has been the lack of an adequate corpus of validated searches that can be utilized as gold standards. Among one-node searches, only two examples have been employed by other groups as gold standards: the Swanson studies of magnesium and migraine [32] and fish oil and Raynaud's phenomenon [33]. We initially felt that it would be impossible to create gold standards in the case of two-node searches, since given a single query (a single pair of literatures), different users might be looking for entirely different types of information. However, in the course of analyzing the two-node searches conducted by field testers, we realized two things: First, once we no longer insisted that LBD searches must predict entirely novel, untested hypotheses, it was relatively easy to ask field testers to score B-terms as relevant or non-relevant. For example, given two literatures concerning specific diseases, we could ask them to identify B-terms that correspond to surgical interventions that are performed in both diseases [2, 3]. Second, we found that B-terms that were marked as useful, interesting or relevant shared certain generic features across many different searches that distinguished them from terms that were marked as non-relevant. Thus, we were able to create manually a corpus of (currently six) diverse gold standard two-node

searches, which have been employed for quantitative modeling [3] and implemented on the Arrowsmith website to rank the terms displayed on the B-list.

As well, we devised a means of creating new gold standard two-node searches and sets of "relevant" B-terms automatically, using a series of 20 templated TREC 2005 Genomics Track queries (http://trec.nist.gov/data/t14_genomics.html) asking for information describing the role(s) of a gene involved in a disease, or describing the role of a gene in a specific biological process. As part of TREC, each query was searched within a biomedical text collection representing a subset of MEDLINE, and a group of judges decided which articles were relevant to the query. We regarded the articles marked as relevant by TREC judges as "gold standards" for each query, and extracted all terms in the titles of these papers. The terms were filtered through a stoplist to remove many of the "uninteresting" terms, and the remaining terms were regarded as capturing some of the known, *explicit* information on each query. Next, we associated each query with a two-node search in which we formulated literature A = the gene name and literature C = the disease or biological process [removing any articles that mention both A and C]. The explicit title terms taken from the gold standard articles in the TREC queries serve the same function for evaluation as does the field-tester marked relevant B-terms in our own six gold standard queries [3]. We suggest that new gold standard searches can be deliberately and perhaps automatically set up for one-node searches as well. For example, in an earlier study of the potential development of viruses as biological weapons, we employed a list of viruses compiled by military experts as a gold standard [34]. One could also follow the lead of one of the Arrowsmith field testers, Ramin Homayouni, who used a set of five genes already known to be part of the reelin signaling pathway as gold standards, and applied a LBD model to a larger list of candidate genes in order to identify genes that are likely to be part of the reelin signaling pathway, even though they do not co-occur in any paper mentioning reelin [35]. This approach makes the admittedly uncertain assumption that the features of known and unknown reelin pathway genes will be similar, but this is a limitation that applies more or less to all gold standards (i.e., the assumption that new instances will be similar to the older ones, as far as their scored features).

From this perspective, it should be an easy process to generate gold standards for one-node searches, as long as two points are kept in mind. First, one must remember that LBD is attempting to model the structure of the scientific literature, not of nature, so making a list derived directly from experimental results, e.g. microarray data, does not suffice to construct a gold standard. Second, one must distinguish LBD systems that make "incremental" predictions from those that attempt to make more radical, cross-disciplinary predictions. Predicting new genes that interact with reelin is an example of the former case. Here, the LBD system merely needs to compare the features associated with a new example against a panel of known positive and negative examples, and identify those that are overall most similar to known examples. In contrast, cross-disciplinary LBD seeks to relate literatures that may appear to have little or nothing in common. The relevant measure is COMPLEMENTARITY, rather than SIMILARITY, since a particular item or concept may link two literatures meaningfully even if it is not prominent in either literature.

Re-defining success in LBD also leads us to re-assess the dichotomy that has been stated as existing between computer-generated and computer-assisted discovery tools. Certainly, we are interested in discoveries that are made by people, not by computers [4] – and yet we have found that B-terms can be automatically ranked in terms of the likelihood that one or more users will find them to be "useful" or "interesting" [3]. Evidently during data mining some nuggets can be seen to be shinier than others, and the computer can present these to the user for further inspection. Actually, the problem is not so much that computers are limited in their ability to predict new discoveries, as that individual scientists vary so widely in their interests and intuitions. It is virtually impossible for any group of scientists to reach consensus in deciding whether a truly novel hypothesis is promising and significant to follow up!

5 Concluding Remarks

If jazz is a sophisticated, intricate form of expression appreciated by the cognescenti, then LBD may be the jazz of informatics. However, jazz enthusiasts probably do not care whether the music they listen to is popular or not, whereas LBD tools were designed for working scientists and our shared goal is to make them both useful and easy to use. Our experiences with the Arrowsmith two-node search have suggested lessons that, we believe, should apply generally to other LBD projects. Most importantly, in LBD, as in jazz, we will best succeed when our different voices and instruments harmonize together.

Acknowledgements The Arrowsmith consortium is supported by NIH, and includes subcontract principal investigators Don Swanson, Maryann Martone, Ramin Homayouni, and Robert Bilder. We also thank Marc Weeber for his assistance and discussions over the years.

References

1. Smalheiser, N. R.: The Arrowsmith project: 2005 status report. In: Conference on Discovery Science 2005. Lecture Notes in Artificial Intelligence, vol. 3735, eds. A. Hoffmann, H. Motoda, and T. Scheffer, Springer, Berlin Heidelberg New York (2005) 26–43
2. Smalheiser, N. R., Torvik, V. I., Bischoff-Grethe, A., Burhans, L. B., Gabriel, M., Homayouni, R., Kashef, A., Martone, M. E., Perkins, G. A., Price, D. L., Talk, A. C., West, R.: Collaborative development of the Arrowsmith two-node search interface designed for laboratory investigators. J. Biomed. Discov. Collab. **1** (2006) 8
3. Torvik, V. I., Smalheiser, N. R.: A quantitative model for linking two disparate sets of articles in Medline. Bioinformatics **23** (2007) 1658–1665
4. Swanson, D. R., Smalheiser, N. R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. **91** (1997) 183–203
5. Bekhuis, T.: Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. Biomed. Digit. Libr. **3** (2006) 2

6. Jensen, L. J., Saric, J., Bork, P.: Literature mining for the biologist: from information retrieval to biological discovery. Nat. Rev. Genet. **7** (2006) 119–129
7. Weeber, M., Kors, J. A., Mons, B.: Online tools to support literature-based discovery in the life sciences. Brief. Bioinform. **6** (2005) 277–286
8. Hristovski, D., Peterlin, B., Mitchell, J. A., Humphrey, S. M.: Using literature-based discovery to identify disease candidate genes. Int. J. Med. Inform. **74** (2005) 289–298
9. Skeels, M. M., Henning, K., Yetisgen-Yildiz, M., Pratt, W.: Interaction Design for Literature-Based Discovery. Proceedings of the ACM International Conference on Human Factors in Computing Systems (CHI 2005), Portland, OR
10. Wren, J. D.: Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics **5** (2004) 145
11. Srinivasan, P., Libbus, B.: Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics **20** Suppl 1 (2004) I290–I296
12. Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., Garner, H. R.: Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics **20** (2004) 389–398
13. Srinivasan, P.: Text Mining: Generating hypotheses from MEDLINE. J. Am. Soc. Inf. Sci. Technol. **55** (2004) 396–413
14. Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R., Molema, G.: Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J. Am. Med. Inform. Assoc. **10** (2003) 252–259
15. Weeber, M., Vos, R., Baayen, R. H.: Using concepts in literature-based discovery: simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. J. Am. Soc. Inf. Sci. Technol. **52** (2001) 548–557
16. Valdes-Perez, R. E.: Principles of human-computer collaboration for knowledge discovery in science. Artif. Intell. **107** (1999) 335–346
17. Kostoff, R. N.: Science and technology innovation. Technovation **19** (1999) 593–604
18. Smalheiser, N. R., Swanson, D. R.: Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput. Methods Programs Biomed. **57** (1998) 149–153
19. Swanson, D. R.: Undiscovered public knowledge. Libr. Q. **56** (1986) 103–118
20. Smalheiser, N. R., Swanson, D. R.: Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neurosci. Res. Commun. **15** (1994) 1–9
21. Smalheiser, N. R., Manev, H., Costa, E.: RNAi and Memory: Was McConnell on the right track after all? Trends Neurosci. **24** (2001) 216–218
22. Smalheiser, N. R.: Predicting emerging technologies with the aid of text-based data mining: the micro approach. Technovation **21** (2001) 689–693
23. Swanson, D. R., Smalheiser, N. R., Bookstein, A.: Information discovery from complementary literatures: categorizing viruses as potential weapons. J. Am. Soc. Inf. Sci. Technol. **52** (2001) 797–812
24. Smalheiser, N. R.: Bath toys: a source of gastrointestinal infection. N. Engl. J. Med. **350** (2003) 521
25. Fuller, S. S., Revere, D., Bugni, P. F., Martin, G. M.: A knowledgebase system to enhance scientific discovery: Telemakus. Biomed. Digit. Libr. **1** (2004) 2
26. Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., Humenik, J. A.: Fullerene data mining using bibliometrics and database tomography. J. Chem. Inf. Comput. Sci. **40** (2000) 19–39
27. Smalheiser, N. R., Torvik, V. I.: Mammalian microRNAs derived from genomic repeats. Trends Genet. **21** (2005) 322–326
28. Smalheiser, N. R., Torvik, V. I.: Alu elements within human mRNAs are probable microRNA targets. Trends Genet. **22** (2006) 532–536
29. Petukhova, O. A., Mittenberg, A. G., Kulichkova, V. A., Kozhukharova, I. V., Ermolaeva, I. uB., Gauze, L. N., Konstantinova, I. M.: A new class of small RNP (alpha-RNP) containing antisense RNA in K-562 cells. IV. The coordinated regulation of the expression of Alu-containing mRNA and alpha-RNA during differentiation. Ontogenez **28** (1997) 437–444

30. Daskalova, E., Baev, V., Rusinov, V., Minkov, I.: Sites and mediators of network miRNA-based regulatory interactions. Evol. Bioinform. Online **2** (2006) 99–116
31. Swanson D. R.: Intervening in the life cycles of scientific knowledge. Libr. Trends **41** (1993) 606–631
32. Swanson D. R.: Migraine and magnesium: eleven neglected connections. Perspect. Biol. Med. **31** (1998) 526–557
33. Swanson D. R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med. **30** (1996) 7–18
34. Swanson, D. R., Smalheiser, N. R., Bookstein, A.: Information discovery from complementary literatures: categorizing viruses as potential weapons. J. Am. Soc. Inf. Sci. Technol. **52** (2001) 797–812
35. Homayouni, R., Heinrich, K., Wei, L., Berry, M. W.: Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics **21** (2005) 104–115