induced apoptotic pathway in the tumor cells. To demonstrate direct telomerase inhibition, it would be crucial to further test whether telomere erosion or eventual growth arrest of cells can be observed by prolonging exposure of the tumor cells to these lectins.

This study represents a major step towards understanding the therapeutic pathways of mistletoe extract treatment in complementary cancer medicine. It has been suggested that the telomerase inhibitors could be better used as adjuvant therapies, in combination with surgery, radiation treatment and chemotherapy with standard agents, because inhibition of the enzyme would require a lag phase before any detrimental effects on the tumor cells. Considering the mistletoe lectin as a telomerase inhibitor, the recommended strategy has been apparently practised for a long time in alternative cancer therapy.

References

1 Kienle, G.S. (1999) The story behind mistletoe: a European remedy from anthroposophical medicine. *Altern. Ther. Health Med.* 5, 34–36
2 Bocci, V. (1993) Mistletoe (*Viscum album*) lectins as cytokine inducers and immunoadjuvant in tumor therapy. A review. *J. Biol. Regul. Homeost. Agents* 7, 1–6
3 Holtskog, R. *et al.* (1988) Characterization of a toxic lectin in Iscador®, a mistletoe preparation with alleged cancerostatic properties. *Oncology* 45, 172–179
4 Valentiner, U. *et al.* (2002) The cytotoxic effect of mistletoe lectins I, II and III on sensitive and multidrug resistant human colon cancer cell lines *in vitro*. *Toxicology* 171, 187–199
5 Büssing, A. *et al.* (1999) Induction of Fas ligand (CD95L) by the toxic mistletoe lectins in human lymphocytes. *Anticancer Res.* 19, 1785–1790
6 Lyu, S.Y. *et al.* (2001) Involvement of caspase-3 in apoptosis induced by *Viscum album* var. *coloratum* agglutinin in HL-60 cells. *Biosci. Biotechnol. Biochem.* 65, 534–541
7 Bantel, H. *et al.* (1999) Mistletoe lectin activates caspase-8/FLICE independently of death receptor signaling and enhances anticancer drug-induced apoptosis. *Cancer Res.* 59, 2083–2090
8 Park, R. *et al.* (2000) Activation of c-Jun N-terminal kinase 1 (JNK1) in mistletoe lectin II-induced apoptosis of human myeloleukemic U937 cells. *Biochem. Pharmacol.* 60, 1685–1691
9 Kim, M.S. *et al.* (2000) Activation of caspase cascades in Korean mistletoe (*Viscum album* var. *coloratum*) lectin-II-induced apoptosis of human myeloleukemic U937 cells. *Gen. Pharmacol.* 34, 349–355
10 Lyu, S.Y. *et al.* (2002) Korean mistletoe lectin-induced apoptosis in hepatocarcinoma cells is associated with inhibition of telomerase via mitochondria controlled pathway independent of *p53*. *Arch. Pharm. Res.* 25, 93–101
11 Levy, M.Z. *et al.* (1992) Telomere end-replication problem and cell aging. *J. Mol. Biol.* 225, 951–960
12 White, L.K. *et al.* (2001) Telomerase inhibitors. *Trends Biotechnol.* 19, 114–120
13 Shay, J.W. *et al.* (2001) Telomerase and cancer. *Hum. Mol. Genet.* 10, 677–685

Shi-Sheng Li
Division of Pharmacognosy
Dept of Medicinal Chemistry
Uppsala University Biomedical Center
Box 574 SE-751
23 Uppsala, Sweden
Present address: University of Virginia,
Dept of Chemistry and Biology
PO Box 400319, Charlottesville
VA 22904-4319, USA

Gaps in text-based knowledge discovery for biology ▼

In the post-genomic era, the emphasis on the use of bioinformatic technology in pharmaceutical research is increasingly shifting from target identification to target ranking and due diligence [1]. New kinds of databases that contain information beyond simple sequences are needed, such as information on subcellular localization, protein interactions, gene regulation and the context of these interactions.

The forerunners of such databases include the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.ad.jp/kegg), the Database of Interacting Proteins (DIP, http://www.dip.doe-mbi.ucla.edu) and the Biomolecular Interaction Network Database (BIND, http://bind.ca). These databases are still small in size and are largely curated by hand. The development of reliable text-based knowledge discovery or literature data-mining technologies can accelerate their growth.

Many example applications of text-based knowledge discovery technologies in biology have been described [2,3]. These examples demonstrated significant progress in terms of both depth and breadth. Text-based knowledge discovery in biology has advanced from the simple recognition of terms to the extraction of interaction relationships from complex sentences. It has also broadened from the recognition of protein interactions to a range of problems, such as improving homology searches, identifying subcellular locations or recognizing themes in the literature. The techniques employed have spanned from word co-occurrence statistics, to pattern matching of linguistic constructs in limited contexts, to powerful natural language processing techniques capable of extracting relations that span multiple sentences through the use of co-reference. These results mark this as an emerging field that provides a synergistic combination of bioinformatics and natural language processing.

Despite the enormous potential for the application of text-based knowledge discovery techniques to biology, few of these techniques have made it into routine use to help manage biological information. We list below some issues that need to be addressed to accelerate the progress and acceptance of this field:

- Abstracts can generally be obtained for free, whereas full papers can generally only be obtained following payment. It is thus tempting to consider applying a literature-mining tool to abstracts. It is crucial to assess, for each type of information that is to be extracted from the literature, whether there is a significant loss if only abstracts are processed, compared with full papers. To date, it appears that no single group has investigated this issue to any great extent.
- Several papers [4,5] focused on extracting the interactions of proteins,

  PII: S1359-6446(02)02415-7

drugs and other molecules from the literature. They variously reported specificity figures from 60–90%. The sensitivity of these systems remains an issue. Moreover, these performance studies were conducted on sample sets that were small in size and also different sample sets were used by different researchers. For an impartial assessment and comparison of the performance of these systems, as well as to understand what works and what does not, it is crucial to do a systematic evaluation. Such an evaluation should be based on a biologically important challenge problem, should have extensive training- and blind-test data, and should have a clear repeatable evaluation metric. To date, this issue seems to have started to catch the attention of researchers – the 'KDD Cup 2002' competitions (held in conjunction with ACM SIGKDD *International Conference on Knowledge Discovery and Datamining*) have included a task on 'Information Extraction from Biomedical Articles', jointly organized by Alex Yeh of MITRE (http://www.mitre.org) and the Flybase group of Harvard (see http://www.biostat.wisc.edu/~craven/kddcup/tasks.html).

- It is reasonable to assume that the completeness and reliability of the outcome of text-based knowledge discovery in biology are dependent on the input documents. Should the selection of input documents be based on keywords, as well as on papers chosen by expert biologists and well-cited articles and their cited references therein, or based on some other methods? To date, it seems that no single group has considered it in this context.
- It is also unclear how well a text-based knowledge discovery system has to perform for it to be useful in biology. To know how good a system has to be, working systems must be given to biologists in user-centered evaluations. To date, it appears that no single group has conducted such a study in any extensive way. We acknowledge, however, that, from experience with previous evaluations in the information retrieval community [6], it is hard to extrapolate from results of batch experiments to predict complex issues of use and user acceptance.

Many such issues have remained unaddressed to date. Nevertheless, text-based knowledge discovery for biology has significant potential, because even imperfect tools are useful if they give improved functionality at low cost.

References

1 Boston Consulting Group (2001) A revolution in R&D – the impact of genomics. *BCG Focus* June, 1–15
2 Hirschman, L. *et al.* Accomplishments and challenges in literature data mining for biology. *Bioinformatics* (in press)
3 Mack, R. and Hehenberger, M. (2002) Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discov. Today*, 7 (Suppl. 11), S89–S98
4 Park, J.C. *et al.* (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Proc. Pac. Symp. Biocomput.* 396–407
5 Putejovsky, J. and Castano, J.M. (2002) Robust relational parsing over biomedical literature: Extracting inhibit relations. *Proc. Pac. Symp. Biocomput.* 362–373
6 Hersh, W. *et al.* (2001) Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Inf. Process. Manag.* 37, 383–402

Limsoon Wong
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613

Cheminformatics – decision making in drug discovery

Roger Lahana, Vice President R&D, Computational Drug Discovery, Synt:em, Parc Scientifique G. Besse, Nimes 30000, France; tel: +33 466 048 668, fax: +33 466 048 667, e-mail: rlahana@syntem.com

Cambridge Healthtech Institute's 6th Annual *Cheminformatics* Conference was one of three conferences grouped together as *Intelligent Drug Discovery and Development 2002* (6–8 May 2002, Philadelphia, PA, USA). The other two conferences were dedicated to lead optimization and screening.

Building smarter libraries

Several aspects of library design were addressed in the first session, chaired by Roger Lahana (Syntem; http://www.syntem.com). Robert Sheridan (Merck Research Laboratories; www.merck.com) shared his experience on searching databases using chemical similarity and described three variations of tricks with chemical similarity that extend the applicability of similarity searching using topological descriptors. The first involves representing mixtures by their descriptor average, which works surprisingly well; the second concerns ways to derive nonpeptide molecules from an

  PII: S1359-6446(02)02368-1