Pacific Symposium on Biocomputing 13:231-242(2008)

SYSTEM-WIDE PERIPHERAL BIOMARKER DISCOVERY USING INFORMATION THEORY[*]

GIL ALTEROVITZ[†]

Division of Health Sciences and Technology, Harvard University/Massachusetts Institute of Technology, Cambridge, MA. Children's Hospital Informatics Program, Boston, MA 02115, USA. Department of Electrical Engineering and Computer Science, Cambridge, MA 02139, USA. Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston 02115, USA.

MICHAEL XIANG[†]

Division of Health Sciences and Technology, Harvard University/Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

JONATHAN LIU

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

AMELIA CHANG

Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston 02115, USA.

MARCO F. RAMONI

Children's Hospital Informatics Program, Boston, MA 02115, USA. Division of Health Sciences and Technology, Harvard University/Massachusetts Institute of Technology, Cambridge, MA. Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston 02115, USA.

The identification of reliable peripheral biomarkers for clinical diagnosis, patient prognosis, and biological functional studies would allow for access to biological information currently available only through invasive methods. Traditional approaches have so far considered aspects of tissues and biofluid markers independently. Here we introduce an information theoretic framework for biomarker discovery, integrating biofluid and tissue information. This allows us to identify tissue information in peripheral biofluids.

[*] This work was supported in part by the National Library of Medicine (NLM/NIH) under grant 5T15LM007092 and the National Human Genome Research Institute (NHGRI/NIH) under grant 1R01HG003354.

[†] These authors contributed equally to the work.

We treat tissue-biofluid interactions as an information channel through functional space using 26 proteomes from 45 different sources to determine quantitatively the correspondence of each biofluid for specific tissues *via* relative entropy calculation of proteomes mapped onto phenotype, function, and drug space. Next, we identify candidate biofluids and biomarkers responsible for functional information transfer ($p < 0.01$). A total of 851 unique candidate biomarkers proxies were identified. The biomarkers were found to be significant functional tissue proxies compared to random proteins ($p < 0.001$). This proxy link is found to be further enhanced by filtering the biofluid proteins to include only significant tissue-biofluid information channels and is further validated by gene expression. Furthermore, many of the candidate biomarkers are novel and have yet to be explored. In addition to characterizing proteins and their interactions with a systemic perspective, our work can be used as a roadmap to guide biomedical investigation, from suggesting biofluids for study to constraining the search for biomarkers. This work has applications in disease screening, diagnosis, and protein function studies.

1. Introduction

The rapidly increasing availability of sequenced genomes since the 1990's has made it clear that genetic analysis alone cannot fully account for organismal complexity[1]. A more complete realization focuses instead on genes' protein products. As such, the field of proteomics aims to understand protein function, structure, and interactions[1].

Proteomics has considerable clinical relevance: proteins carry out cellular functions, comprise drug targets, and often participate in or indicate disease pathogenesis. For example, a doctor may take a blood sample to perform a liver function test[2], for which certain enzymes (*e.g.*, alanine transaminase) are elevated in liver dysfunction. Recently, biomarkers for various diseases have emerged, including prostate specific antigen (PSA) for prostate cancer[3] and C-reactive protein (CRP) for heart disease[4]. Therefore, identification of clinically significant protein biomarkers of phenotype and biological function is an exciting and expanding area of research that promises to extend diagnostic capabilities.

The use of biomarkers from easily accessible biofluids (*e.g.* blood, urine) is advantageous for evaluating the state of harder-to-reach tissues and organs. Biofluids capture proteins and protein fragments released by cells in the body, either as waste or to communicate with other cells or tissues[5]. In addition, biofluids are much more readily accessible, unlike more invasive or unfeasible techniques such as tissue biopsies (e.g. brain tissue). To date, however, approaches to biomarker prediction have analyzed tissues and biofluids separately[6].

Here we propose an information theoretic framework for discovery of novel biomarkers that utilizes information from biofluid proteins that can serve as functional, phenotypic, and drug interaction proxies for the underlying tissues. In order to specify a biomarker, a researcher must identify both a biofluid (*e.g.* blood) and protein(s) in that biofluid that are relevant. Due to the presence of dozens of biofluids and many thousands of proteins, too many combinatorial possibilities exist for them to be tested individually. In this work we propose methods to identify both biofluids and specific proteins that are particularly well suited for biomarker discovery and validation.

Biofluids contain proteins from tissues and serve as effective communication/hormonal . Conceptually, the tissue acts as a transmitter of information and the biofluid (sampled by the physician) as a receiver. The informativeness of the biofluid is reliant on the fidelity of the channel. Sources of noise which decrease fidelity include addition of proteins derived from other tissues or from the biofluid itself; proteins may also be lost through the glomerular filtration process that removes proteins smaller than 45 kDa from plasma[7]. These factors can substantially bias the protein composition of a biofluid: for instance, the plasma abundances of interleukin-6 and albumin differ[8] by 10 orders of magnitude. Additionally, looking simply at protein overlap would miss information transmission that occurs through classes of proteins and protein-protein interactions. Thus, we consider not the proteins directly, but instead their projection onto functional, drug, and disease spaces, allowing the measurement of functional distance between tissues and biofluids. Closeness in these abstract spaces signifies a low level of distortion across the information channel, and hence high informativeness of the biofluid.

It turns out that information theory has already developed a robust, principled framework[9] for evaluating such a channel problem. The informativeness of a biofluid for a tissue can thus be evaluated within this framework and be used to guide disease and physiological investigations.

2. Methods

In total, 26 human proteomes[10] were obtained from 45 studies. The 16 tissue proteomes comprised brain, cartilage, cornea, heart, kidney, larynx, liver, macrophage, muscle, nose, ovary, pancreas, pituitary, platelet, skin, and stomach. The 10 biofluid proteomes comprised amniotic fluid, cerebrospinal fluid, plasma, pleural fluid, saliva, serum, sputum, synovial fluid, tear, and urine. The full human proteome was obtained from the Gene Ontology Annotation database[11]. The Gene Ontology[12], or GO (23,692 terms) was used to map proteins to functional space, employing the three hierarchies of cellular component, biological process, and molecular function. A controlled vocabulary for diseases (5,648 terms) was extracted from Online Mendelian Inheritance in

Man (OMIM)[13] to construct the disease-based ontology for mapping proteins to disease space. Similarly, the drug-based ontology (411 terms) was created using the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB)[14] for mapping proteins to drug space. Of the three ontologies, GO has the largest number of terms and represents the most comprehensive distribution of information. As such, although all three ontologies were used to identify significant tissue-biofluid relationships, focus was on the GO-derived results in the identification of candidate biomarkers.

In information theory, relative entropy[15] is a measure of the distance between an unobserved distribution T (here: tissue) and an observed distribution B (here: biofluid). Lower relative entropy denotes closer correspondence between the two distributions. The relative entropy R between a tissue T and a biofluid B was determined as:

$$R(B,T) = \sum_{n=1}^{N} b(V_n) \log \frac{b(V_n)}{t(V_n)}$$

Here, $b(V_n)$ and $t(V_n)$ denote the annotation frequency of term V_n across B and T, respectively, and N is the total number of terms in the function, disease, and drug space. For example, if B = urine, with 1000 proteins, and V_n = “ion binding”, then $b(V_n) = 0.01$ means that 1% (10) proteins in the urine proteome are associated with “ion binding” function. A total of 36,000 relative entropy simulations were performed between the tissues and randomly-chosen sets of proteins from the entire proteome to ascertain the significance of tissue-biofluid connections, followed by application of Bonferroni multiple test correction[16]. Thus, a biofluid B is informative of a tissue T if its relative entropy score $R(B,T)$ with that particular tissue is significantly better than the relative entropy scores of randomly-chosen protein sets with the same number of proteins as B (i.e., $p < 0.01$, after multiple test correction). The approach we proposed is diagramed in Figure 1. Connections were considered significant only if they were significantly better than random in terms of channel information faithfulness for all three spaces: function, disease, and drug.

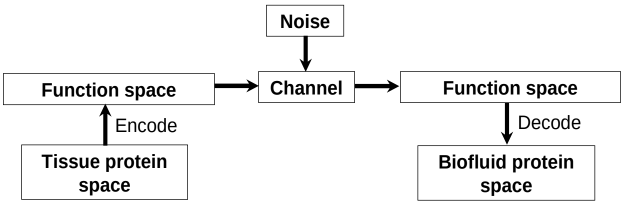

Figure 1. Information theoretic characterization of biofluid-tissue interaction

A search was done for all GO terms with representation at significantly similar frequencies between *B* and *T*. Such terms were termed "bandwidth-carrying" terms because they are primarily responsible for transfer of functional information across the tissue-biofluid channel (*i.e.*, based on relative entropy score; see above). Fisher's exact test was used to compute the probability $p(V_n, B, T \mid F)$ of selecting, from the full human proteome *F*, a random protein sample the same size as *B* sharing, with *T*, the same level of frequency similarity or better for V_n:

$$p(V_n, B, T \mid F) = \sum_{k=i}^{j} \frac{\binom{|F(V_n)|}{k}\binom{|F| - |F(V_n)|}{|B| - k}}{\binom{|F|}{|B|}}$$

Here, $F(V_n)$ constitutes all human proteins annotated by V_n, and set size notation is used. *k* ranges between $i = |B| \cdot b(V_n)$ and $j = |B| \cdot [2t(V_n) - b(V_n)]$, denoting counts of equal or higher similarity for V_n. For example, say *T* and *B* both contain 100 proteins; 40 proteins (40%) in tissue *T* are annotated by V_n = "ion binding" (Gene Ontology), and 38 proteins (38%) in biofluid *B* are annotated by V_n. In this case, $i = 38$ and $j = 42$ for an equal or more similar level of V_n frequency (between 38% and 42%; *i.e.* within 2% of 40%) with *T* as the specific biofluid *B*.

Candidate biomarkers were selected using a scoring process. For a given tissue-biofluid combination, proteins were scored by summing the Shannon information content[17] of the tissue-biofluid pair's "bandwidth-carrying" terms residing in the protein's list of GO annotations. Candidate biomarkers were chosen as biofluid proteins with $p < 0.05$ compared to the scores of randomly-chosen proteins from the full human proteome.

3. Results

3.1 Significant Tissue-Biofluid Channels

A total of 9 biofluids were found to be significantly informative for a total of 14 tissues. In all, 26 tissue-biofluid channels were significant with $p < 0.01$ after Bonferroni correction, while 10 additional tissue-biofluid channels had borderline significance of $p < 0.05$. Figure 2 displays significant channels ($p < 0.01$) between tissues (rectangles) and biofluids (circles). Two tissues, the heart and the pituitary, were not found to be significant with any biofluid tested. On the other hand, the tear biofluid was not found to be informative for any tissue.

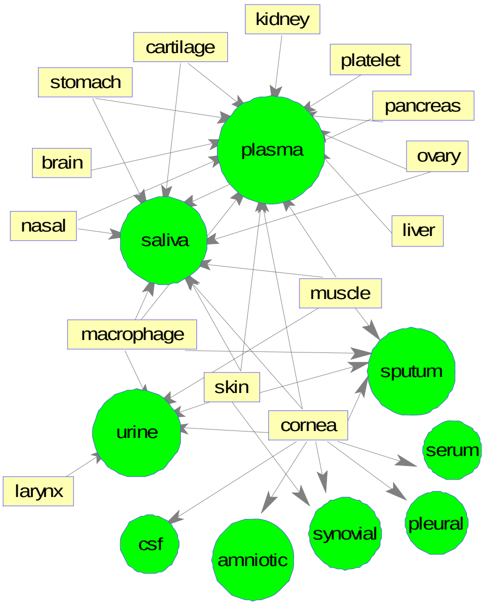

Figure 2. Significant tissue-biofluid channels

As might be expected, blood plasma was significantly informative of most tissues (exceptions were heart, larynx, and pituitary). Interestingly, saliva was the next most informative biofluid (significance across 9 tissues), followed by sputum and urine, which were informative of 4 and 5 tissues respectively. The remaining biofluids (except tear) were informative of 1-2 tissues each.

Corneal tissue shared significance with many biofluids, including cerebrospinal fluid (CSF). This connection has been noted in the literature; for instance, one study noted elevated insulin concentrations in the cornea and CSF upon intralumbar injection[18]. Moreover, topical application of insulin eye drops caused its accumulation within CSF[19]. Sclerotic diffusion could account for the detection of inflammatory response in CSF upon corneal inoculation of herpes simplex virus[20]. Interestingly, CSF was not found to be significantly informative for brain tissue, perhaps because CSF only interacts with the outer edge of the brain. Indeed, one common use of CSF is to diagnose meningitis, which is an infection of the membrane that covers the brain and not of the brain itself.

Significant connections found between the cornea and the other biofluids have also been cited in previous studies. For example, the cornea has been associated with synovial fluid; arthritis patients often display upregulation of proinflammatory cytokines in synovial fluid and corneal samples[21]. Additionally, identical bacteria can be isolated from cornea and sputum during nosocomial eye infection[22]. Literature corroboration increases confidence in our method; consequently, associations with other biofluids that have not been thoroughly explored to date can serve as useful avenues for investigation.

Another significant relationship was discovered between macrophage and sputum, which can be rationalized by macrophages' role in the removal of necrotic debris from the lungs. This tissue-biofluid link is supported by studies showing that increased inflammatory cytokine levels in sputum stimulate macrophage production of metalloproteinases[23]. Other studies have used induced sputum to determine macrophage phenotypes in airway afflictions[24]. Since macrophages are highly involved in immune diseases due to their phagocytic capacity, further elucidation of this relationship could have a variety of clinical applications.

3.2 Identification of Candidate Biomarkers

To identify actual candidate biomarkers, we discovered "bandwidth-carrying" GO terms responsible for transmitting the bulk of functional information from tissue to biofluid. Note that such "bandwidth-carrying" terms can exist between a tissue and biofluid even when the overall biofluid was not found to be informative of the overall tissue. Between the 16 tissues and the 10 biofluids, 519 "bandwidth-carrying" terms were identified with $p < 0.001$. Using these terms, 851 unique proteins were identified as candidate biomarkers for the 16 tissues in this study. Plasma was the most productive biofluid, containing an average of 269 candidate biomarkers per tissue; serum was next with an average of 112 biomarkers per tissue. Other biofluids presented varying numbers of candidate biomarkers: urine for instance had an average of 37 biomarkers per tissue, whereas tear was not found to contain any candidate biomarkers for any tissue with the sole exception of cornea. A portion of the resulting network (e.g. for ovary and biofluids) is shown in Figure 3.

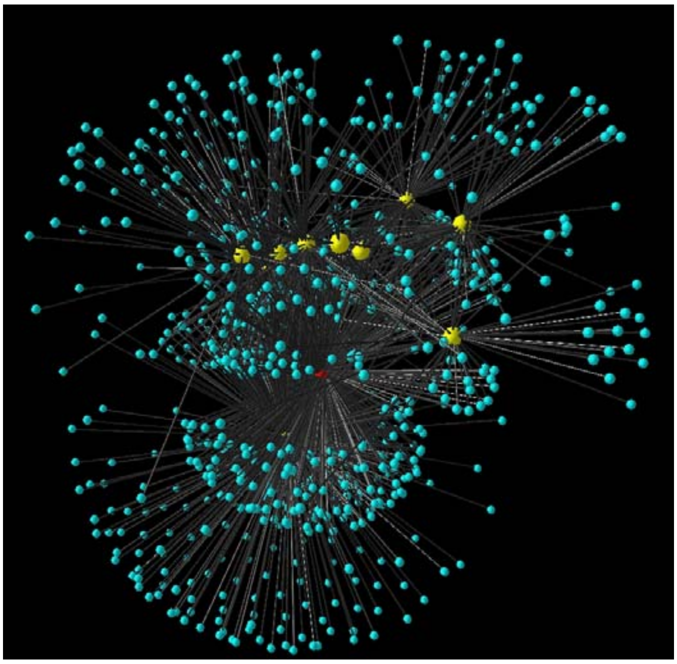

Figure 3. A portion of the tissue-biomarker-biofluid network. Ovary (large, dark sphere in center), biofluids (small spheres), and candidate protein biomarkers (large, light spheres).

Since our approach assesses function on a tissue-wide level, our candidate biomarkers are not restricted to any particular cellular or pathological process. However, their contributions to physiological state can be hypothesized from their functional annotations, and hence serve as initial candidates for screening. A quick scan of the candidate biomarkers reveals some proteins that have been discovered by traditional means. For example, in our list of potential biomarkers for measuring ovarian function, we found a number of known cancer markers. Represented in this list were epidermal growth factor receptor (EGFR), BRCA1, and Apolipoprotein E. These proteins are clinically significant markers of ovarian cancer: EGFR is a specific target for ovarian cancer therapy[25], and mutation of BRCA1 correlates with ovarian cancer risk[26]. Apolipoprotein E has been found to be upregulated in ovarian cancer[27] and also critical for cell survival and proliferation in the disease[28].

Although these ovarian cancer biomarkers have already been validated, the need for additional biomarkers is striking. Half of ovarian cancer patients initially present at Stage III or Stage IV when 5-year survival is only 20%, thus making the disease responsible for more deaths than all other gynecological cancers combined[29]. The successful identification of known ovarian cancer biomarkers confirms our approach, suggesting that the list of predicted ovary biomarkers contains promising targets for clinical investigation. The rapid, guided testing of novel biomarkers can then improve the understanding and treatment of ovarian cancer.

3.3 Establishing Biomarker Quality

The overall quality of the candidate biomarkers was assessed by measuring co-citation frequencies in PubMed (Figure 4). The co-citation frequency of each biomarker with the corresponding predicted target tissue(s) ("Predicted" bar) was compared with that of the same biomarkers but with non-corresponding (off-target) tissues, for which the biomarker was not predicted to be informative ("Non-predicted" bar). Results were then sampled for manual verification of the links within the papers as well as the underlying Medical Subject Headings (MeSH).

The median number of publications co-citing a given tissue and one of its predicted biomarkers was 24, while the median number for non-predicted biomarker/tissue combinations was 16. This difference was found to be significant by the Mann-Whitney *U*-test ($p < 5.3x10^{-13}$).

Tissue specificity, and hence confidence in biomarker quality, was improved further by filtering the candidate biomarker list according to the significant tissue-biofluid channels (see section 3.1 and Figure 2). Thus, a candidate biomarker was considered only if the biofluid containing the biomarker was found to be significantly informative of the biomarker's target tissue. The filtered list, comprising 519 unique biomarkers, was about 60% of the size of the unfiltered list. However, the filtered biomarkers were even more tissue-specific: the median co-citation rate of predicted biomarkers/tissues was 28 publications, whereas the median co-citation rate of non-predicted biomarkers/tissues remained at 16 publications ($p < 5.5x10^{-21}$). This increase after filtering (Figure 4) suggests that clinically relevant protein biomarkers of a tissue are likely to reside in the biofluids found to be significantly informative of that tissue, thus integrating the information channel model of tissue-biofluid interaction (*via* relative entropy) with biomarker prediction and discovery.

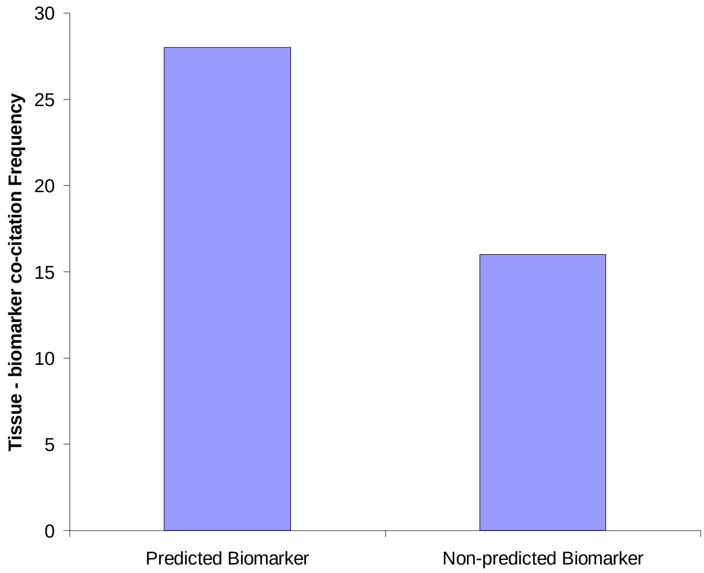

Figure 4. Tissue specificity of candidate biomarkers.

An example case study is Amyloid beta A4 protein. It was found to be modulated with ovariectomies in previous studies[30], thus helping validate the potential of evaluating ovarian function through such a biomarker. On the other hand, it was not found to been previously associated with ovarian cancer in the literature. As an indicator of function, this biomarker can be potentially informative about phenotypic state as well. To validate this, we analyzed an independently performed gene expression study of ovarian cancer[31] and found this protein was significantly upregulated ($p < 0.01$). To control the false discovery rate, we calculated the q-value[32] as 1.4×10^{-10}.

By combining protein interactions, gene expression, and PubMed with an information theoretic framework, this approach promises to allow for the discovery of novel functional and phenotypic biomarkers of internal tissue processes.

4. Discussion and Conclusion

Our framework combines biofluid and tissue information for the discovery of novel biomarkers. Unlike prior work, our approach takes advantage of functional synergy between certain biofluids and tissues with the potential for clinically significant findings not possible if tissues and biofluids were considered individually.

By conceptualizing tissue-biofluid interactions as information channels, we identified significant biofluid proxies that can be used for guided development of clinical diagnostics. We then predicted candidate biomarkers based on information transfer criteria across the tissue-biofluid channels. Significant biofluid-tissue relationships can be used to prioritize clinical validation of new biomarkers.

Some of our results have already been validated for clinical utility, increasing confidence in our findings. At the same time, many are currently novel, suggesting that multiple additional biomarkers can be experimentally confirmed regarding clinical significance. Our work provides a new approach for linking molecular bioinformatics to clinical research, with the potential to expand physiological, phenotypic, and clinical diagnostic capabilities for applications in biology and medicine.

References

[1] M. Tyers and M. Mann, *Nature* **422** (6928), 193 (2003).

[2] S. G. Sakka, *Current opinion in critical care* **13** (2), 207 (2007).

[3] E. A. Singer, D. F. Penson, and G. S. Palapattu, *Jama* **297** (9), 949; author reply 949 (2007).

[4] D. C. Crawford, C. L. Sanders, X. Qin et al., *Circulation* **114** (23), 2458 (2006).

[5] L. A. Liotta, M. Ferrari, and E. Petricoin, *Nature* **425** (6961), 905 (2003).

[6] Y. D. He, *Cancer Biomark* **2** (3-4), 103 (2006).

[7] J. M. Jacobs, J. N. Adkins, W. J. Qian et al., *J Proteome Res* **4** (4), 1073 (2005).

[8] N. L. Anderson and N. G. Anderson, *Mol Cell Proteomics* **1** (11), 845 (2002).

[9] Gil Alterovitz, Michael Xiang, and Marco F. Ramoni, presented at the Proceedings of the Information Theory Applications Workshop, San Diego, CA, 2007 (unpublished).

[10] S. Hu, J. A. Loo, and D. T. Wong, *Proteomics* **6** (23), 6326 (2006); M. A. Tangrea, B. S. Wallis, J. W. Gillespie et al., *Expert review of proteomics* **1** (2), 185 (2004).

[11] E. Camon, M. Magrane, D. Barrell et al., *Nucleic Acids Res* **32** (Database issue), D262 (2004).

[12] M. Ashburner, C. A. Ball, J. A. Blake et al., *Nature genetics* **25** (1), 25 (2000).

[13] D. L. Wheeler, T. Barrett, D. A. Benson et al., *Nucleic Acids Res* **34** (Database issue), D173 (2006).

[14] M. Hewett, D. E. Oliver, D. L. Rubin et al., *Nucleic Acids Res* **30** (1), 163 (2002).

[15] David J. C. MacKay, *Information theory, inference, and learning algorithms*. (Cambridge University Press, Cambridge, U.K. ; New York, 2003).

[16] Martin Bland, *An Introduction to Medical Statistics*, 3rd ed. (Oxford University Press, 2000).

[17] G. Alterovitz, M. Xiang, M. Mohan et al., *Nucleic acids research* **35** (Database issue), D322 (2007).

[18] S. B. Koevary, V. Lam, and G. Patsiopoulos, *Optometry* **75** (3), 183 (2004).

[19] S. B. Koevary, V. Lam, G. Patsiopoulos et al., *J Ocul Pharmacol Ther* **19** (4), 377 (2003).

[20] R. H. Boerman, A. C. Peters, B. R. Bloem et al., *Acta Neuropathol (Berl)* **83** (3), 300 (1992).

[21] J. Prada, B. Noelle, H. Baatz et al., *Br J Ophthalmol* **87** (5), 548 (2003).

[22] E. Hilton, A. A. Adams, A. Uliss et al., *Lancet* **1** (8337), 1318 (1983).

[23] K. F. Chung, *Curr Drug Targets* **7** (6), 675 (2006).

[24] J. Domagala-Kulawik, M. Maskey-Warzechowska, J. Hermanowicz-Salamon et al., *J Physiol Pharmacol* **57 Suppl 4**, 75 (2006).

[25] H. Lassus, H. Sihto, A. Leminen et al., *Journal of molecular medicine (Berlin, Germany)* **84** (8), 671 (2006).

[26] M. C. King, J. H. Marks, and J. B. Mandell, *Science* **302** (5645), 643 (2003).

[27] C. D. Hough, K. R. Cho, A. B. Zonderman et al., *Cancer research* **61** (10), 3869 (2001).

[28] Y. C. Chen, G. Pohl, T. L. Wang et al., *Cancer research* **65** (1), 331 (2005).

[29] R. Gogoi, S. Srinivasan, and D. A. Fishman, *Expert review of molecular diagnostics* **6** (4), 627 (2006).

[30] S. S. Petanceska, V. Nagy, D. Frail et al., *Experimental gerontology* **35** (9-10), 1317 (2000).

[31] D. Roberts, J. Schick, S. Conway et al., *British journal of cancer* **92** (6), 1149 (2005).

[32] J. D. Storey and R. Tibshirani, *Proceedings of the National Academy of Sciences of the United States of America* **100** (16), 9440 (2003).

[33] T. I. Williams, K. L. Toups, D. A. Saggese et al., *Journal of proteome research* (2007).