Combining Semantic Relations and DNA Microarray Data for Novel Hypotheses Generation

Dimitar Hristovski, PhD,[1] Andrej Kastrin,[2] Borut Peterlin, MD[2]

and Thomas C. Rindflesch, PhD[3]

[1]Institute of Biomedical Informatics, Faculty of Medicine, Ljubljana, Slovenia
[2]Institute of Medical Genetics, University Medical Centre, Ljubljana, Slovenia
[3]National Library of Medicine, NIH, Bethesda, MD, USA

dimitar.hristovski@mf.uni-lj.si, {andrej.kastrin, borut.peterlin}@guest.arnes.si, tcr@nlm.nih.gov

Abstract. Although microarray experiments have great potential to support progress in biomedical research, results are not easy to interpret. Information about the functions and relations of relevant genes needs to be extracted from the vast biomedical literature. A potential solution is to use computerized text analysis methods. Our proposal enhances these methods with semantic relations. We describe an application that integrates such relations with microarray results and discuss its benefits in supporting enhanced access to the relevant literature for interpretation of results and novel hypotheses generation. The application is available at http://sembt.mf.uni-lj.si.

Keywords: microarray analysis; literature-based discovery; semantic predications; natural language processing

1 Introduction

Microarray technology can be used to measure the expression levels of essentially all genes within a genome and can provide insight into gene functions and transcriptional networks [1]. This wealth of information potentially underpins significant advances in biomedical knowledge. However, successful use of microarray data is impossible without comparison to published documents. Due to the large size of the life sciences literature, sophisticated information management techniques are needed to help assimilate online textual resources.

Automatic text mining, commonly based on term co-occurrence, has been used to identify information valuable for interpreting microarray results. In this paper we propose the use of semantic relations (or predications) as a way of extending these techniques. Semantic predications convert textual content into "executable

knowledge" amenable to further computation supporting research on genes and relevant diseases. In addition, we suggest that the combination of microarray data and semantic predications can profitably be exploited in the literature-based discovery (LBD) paradigm to further enhance the scientific process.

We describe the use of SemRep [2] for extracting a wide range of semantic predications from MEDLINE citations and discuss a tool for manipulating a database of such relations. We then exploit these predications and the results of a microarray experiment from the GEO repository (GSE8397) [3] on Parkinson disease to generate novel hypotheses in the LBD paradigm.

2 Background

A variety of statistical techniques have been used to manipulate text features (usually in MEDLINE citations) to elucidate relevant literature on microarray experiments. Shatkay et al. [4], for example, extract gene function terms from a set of citations identified as related to a kernel document using a document similarity algorithm. Many methods use co-occurring text words [5], often in conjunction with additional information such as MeSH indexing or structured information from related databases such as the Gene Ontology (e.g. [6, 7]). Some systems exploit a thesaurus to identify concepts in text [8] or calculate implicit information by identifying terms related through co-occurrence with shared, intermediate terms [9].

The LBD paradigm was introduced by Swanson [10] for discovering new relations (hypotheses) between concepts by analyzing the research literature. Swanson's method and most of those that followed, including our BITOLA system [11], are co-occurrence based. We expanded the LBD paradigm by using semantic relations and discovery patterns [12], and we applied the expanded methodology to investigate drug mechanisms [13]. In this paper we further expand LBD by combining microarray data with semantic relations extracted from the literature and by defining new discovery patterns.

The SemRep program extracts semantic predications from MEDLINE citations in several domains, including clinical medicine [2], molecular genetics [14], and pharmacogenomics [15]. The system is symbolic and rule based, relying on structured domain knowledge in the Unified Medical Language System® (UMLS),® extended for molecular genetics and pharmacogenomics. SemRep uses underspecified syntactic analysis, in which only simple noun phrases are identified. MetaMap is used [16] to identify Metathesaurus concepts and is augmented by ABGene [17] to identify gene names. Text tokens marked as potential gene names by either MetaMap or ABGene are searched in a precomputed Berkeley DB table compiled from Entrez Gene official symbols, names, aliases, and identifiers. A successful match is given the Entrez Gene identifier. The gene table is updated periodically and is currently limited to human genes.

SemRep predications have Metathesaurus concepts as arguments and Semantic Network relations as predicates. The relations currently addressed are:

Genetic Etiology: ASSOCIATED_WITH, PREDISPOSES, CAUSES

Substance Relations: INTERACTS_WITH, INHIBITS, STIMULATES

Pharmacological Effects: AFFECTS, DISRUPTS, AUGMENTS

Clinical Actions: ADMINISTERED_TO, MANIFESTATION_OF, TREATS

Organism Characteristics: LOCATION_OF, PART_OF, PROCESS_OF

Co-existence: CO-EXISTS_WITH

As an example, SemRep extracts the predication "MDB1 CAUSES Autistic Disorder" from the text … *the loss of Mbd1 could lead to autism-like behavioral phenotypes* … In this interpretation, Mbd1 has semantic type 'Gene or Genome' and *autism* maps to the concept "Autistic Disorder" (with semantic type 'Disease or Syndrome'). *Lead to* is an indicator for the semantic relation CAUSES. Similarly the predication "MBD1 INTERACTS_WITH HTR2C" is extracted from … *Mbd1 can directly regulate the expression of Htr2c, one of the serotonin receptors,* … on the basis of the identification of the two genes in this text and the verb *regulate* indicating the relation INTERACTS_WITH.

3 Methods

We processed microarray data from the GEO data set and integrated it with SemRep predications in a MySQL database. To accommodate literature-based discovery we formulated discovery patterns [12] that refer to the interaction of drugs and genes. Finally, we devised tools for searching the database using the discovery patterns in order to explore the microarray data and associated research literature for a specific disease and suggest hypotheses about potential drug therapies for that disease.

3.1 Preparing the microarray experiments and results

Currently, we have preprocessed and integrated only a few microarray datasets. In the future we will consider allowing the user to request any GEO dataset be processed with default parameters. Another option is to allow the user to upload a list of differentially expressed genes directly into the system. Below we describe the processing of a GEO dataset that is used throughout the paper to illustrate our methodology and the tools.

A total of 47 Affymetrix HG-U133A CEL files for 29 Parkinson disease patients and 18 controls were retrieved from the GEO repository (GSE8397) [3]. All computations were carried out in the R software environment for statistical computing using additional Bioconductor packages [18, 19]. The normalization of the raw data was performed using the MAS5 algorithm as implemented in the `affy` package. Hybridization probes were mapped to Entrez Gene IDs by annotation data in the `hgu133a.db` package. Analysis of differentially expressed genes (DEG) was performed using Welch's t-test from the `multtest` package. The Benjamini and Hochberg method was selected to adjust *p*-values for multiple testing [20]. As a

confidence threshold we used an adjusted value of $p \leq 0.01$. A total of 567 DEGs were used for further processing.

3.2 Integrated database with semantic relations and microarray results

We built an integrated MySQL database to store the semantic relations extracted by SemRep and the microarray results we processed. The data is spread across several tables holding information on the arguments and relations from the predications. For each argument we store concepts and synonyms as well as semantic types. Arguments are UMLS concepts, but when an argument is a gene, in addition to the UMLS CUI (Concept Unique Identifier) we also store the Entrez Gene ID, which serves as a link to the microarray results. In addition, a link is maintained to the sentence in the MEDLINE citation from which the predication was generated.

We have developed two tools for searching the integrated database: one for searching direct relations between concepts and one for indirect relations. In both cases the arguments of the relations can be limited to genes from the microarray. To allow fast and flexible searching of the integrated database we use Lucene and have built separate indexes, one for fast text searching with Lucene and another for accessing the data stored in MySQL when needed. The tools for searching are Web based and were built with the Ruby on Rails application development framework. The tools provide a flexible way to answer questions about what is already known from the literature: genes associated with a disease; relations between a disease and other concepts; relations between the genes from the microarray and themselves or with other concepts. The tools can also generate novel hypotheses: implicit links between a disease and up- or downregulated genes; concepts that might be used to affect these genes; and potential new treatments.

3.3 Discovery patterns for novel hypotheses generation

For novel hypotheses generation, the tools exploit *discovery patterns,* which are query combinations whose results represent a novel hypothesis – something not specified in the literature or in the microarray results alone. We have designed several new discovery patterns, only two of which are described here. The two discovery patterns, which can be used to discover new therapeutic approaches for some disease, work by regulating the up- or downregulated genes related to that disease (Figure 1).

For example, if we want to investigate regulating genes that are upregulated in the microarray, we search for concepts (genes, drugs, etc.) that are reported in the literature as inhibiting the upregulated genes. We call this discovery pattern "inhibit the upregulated." Similarly, we can investigate downregulated genes with the "stimulate the downregulated" pattern, in which case we search for biomedical concepts that are known to stimulate the downregulated genes.

These discovery patterns combine information from the microarray data about up- or downregulated genes in patients having a certain disease with information from the literature about biomedical concept that can be used to regulate those genes. The discovery patterns can be more complex and involve the combination of more

searches through several common intermediate concepts. Also, relations in addition to "INHIBITS" and "STIMULATES," could be used. The novel hypotheses produced by discovery patterns need to be evaluated by a human expert, first by reading the literature and then by laboratory experiments.

Our tools allow complex queries implementing discovery patterns to be specified easily. As output, semantic relations or novel hypotheses are presented first. Then, on request, the highlighted sentences and MEDLINE citations from which the semantic relations are extracted are shown. Some examples are given in the next section.

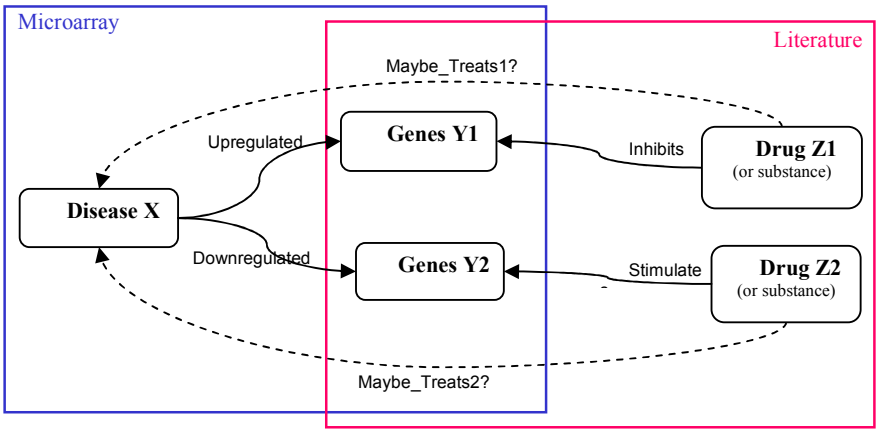


Fig. 1. The two discovery patterns "Inhibit the Upregulated" and "Stimulate the Downregulated" that can be used to find novel therapeutic agents. The patterns combine information from the microarray about which genes are up- or downregulated with information from the literature about which agents could be used to inhibit or stimulate these genes.

4. Results

4.1 Numbers describing size of processing

We used SemRep to process 43,369,616 sentences from 6,699,763 MEDLINE citations published between 1999 and the end of March 2009. 21,089,124 semantic predication instances were extracted, representing 7,051,240 distinct predication types. There are 1,334,014 distinct UMLS concepts appearing as arguments of the semantic predications.

4.2 Evaluation

Evaluating medical aspects of our hypotheses is beyond the scope of this work, and in this paper we do not address the reliability of microarray results. Our focus is on estimating SemRep accuracy, and for this we rely on the work of Masseroli et al [14]. They established a baseline by calculating precision on 2,042 relations extracted with SemGen (now integrated into SemRep): 41.95% for 'genetic' relations (INHIBITS, STIMULATES) and 74.2% for 'etiologic' relations (CAUSES, ASSOCIATED_WITH, PREDISPOSES). They then propose a postprocessing strategy to improve results using the distance (measured in phrases) of the argument (subject and object) from the indicator of the semantic relation. For example, if INHIBITS and STIMULATES relations are filtered for arguments at distance 1 from the indicator, precision increases to 70.75%; however, recall drops to 43.6%. At argument distance 2 (or less) from the indicator, precision is 55.88% and recall 66.28%. In exploiting this method, we first show the user relations more likely to be correct by ranking results in order of increasing argument-predicate distance.

4.3 Generating novel hypotheses for potential therapeutic agents

We illustrate the capabilities of our methodology on a microarray for Parkinson disease (PD) (GEO GSE8397) [21] and investigate therapies which might inhibit the expression of upregulated genes or stimulate the expression of downregulated genes associated with this disorder.

4.3.1 Inhibit the upregulated

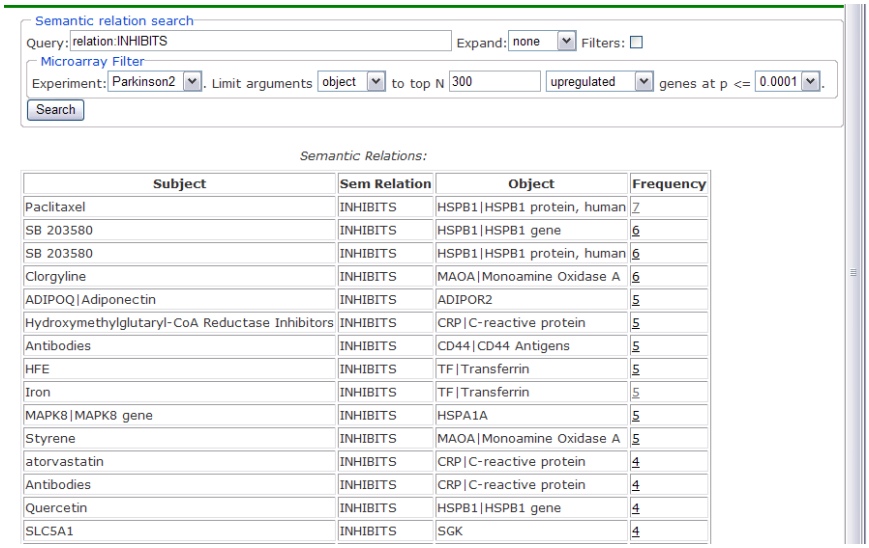
Figure 2 shows how the "inhibit the upregulated" pattern is implemented with our tool for searching direct semantic relations from the literature. In the *Query* field we can enter a simple or more complex Boolean query expression. The query terms, by using an appropriate short field name, can refer to the name, semantic type and concept identifier (UMLS CUI or Entrez Gene ID) of the subject and/or object of the semantic relation as well as the name of the relation. In Figure 2 we entered "relation:INHIBITS" which means we want to search for all the biomedical concepts where one of them "INHIBITS" the other. If we select the *Search* button without providing additional constraints we will get the first 20 of about 300,000 "INHIBITS" relations.

To completely implement the "inhibit the upregulated" discovery pattern, we provide an additional constraint in the "Microarray Filter" group of fields. The first field *Experiment* allows us to select the microarray experiment (the default value is *none*). In our case we select a PD experiment denoted here as *Parkinson2* (corresponds to GEO GSE8397). The next field, *Limit arguments,* allows us to select which argument of the semantic relation we want to limit. We have selected *object,* which means that the object of the "INHIBITS" relation must be one of the genes on the selected microarray. The other possibilities for this field are: *any,* meaning we are interested in relations where at least one of the arguments is a gene from the

microarray; *subject,* meaning the subject of the relations has to be one of the microarray genes; and *both,* meaning only direct relations between the genes on the microarray are to be retrieved.

The next two fields allow us to specify the number of microarray genes to be used for filtering. Here *top N* refers to the most differentially expressed genes. We can select only the *upregulated* or the *downregulated* or *any,* meaning the top N up- or downregulated. Because of performance and implementation issues the top N currently can not be more than 400 genes. The final field in this group allows us to select genes based on the p value. The upper part of Figure 2 shows the options specified for the following example.

As a result of the query we get a list of semantic relations ordered by ascending frequency. For each relation, the subject, the relation itself, the object, and frequency of occurrence are shown. Frequency of occurrence indicates the number of sentences from which the semantic relation was extracted. The frequency number is actually a hyperlink which can be selected to show the list of sentences from which the relations were extracted (subject, relation, and object are highlighted). Additionally, the PubMed ID (PMID) is provided for each sentence; this can be selected to show the PubMed citation in which the sentence appears. Examples of this are shown below.



Subject	Sem Relation	Object	Frequency
Paclitaxel	INHIBITS	HSPB1|HSPB1 protein, human	7
SB 203580	INHIBITS	HSPB1|HSPB1 gene	6
SB 203580	INHIBITS	HSPB1|HSPB1 protein, human	6
Clorgyline	INHIBITS	MAOA|Monoamine Oxidase A	6
ADIPOQ|Adiponectin	INHIBITS	ADIPOR2	5
Hydroxymethylglutaryl-CoA Reductase Inhibitors	INHIBITS	CRP|C-reactive protein	5
Antibodies	INHIBITS	CD44|CD44 Antigens	5
HFE	INHIBITS	TF|Transferrin	5
Iron	INHIBITS	TF|Transferrin	5
MAPK8|MAPK8 gene	INHIBITS	HSPA1A	5
Styrene	INHIBITS	MAOA|Monoamine Oxidase A	5
atorvastatin	INHIBITS	CRP|C-reactive protein	4
Antibodies	INHIBITS	CRP|C-reactive protein	4
Quercetin	INHIBITS	HSPB1|HSPB1 gene	4
SLC5A1	INHIBITS	SGK	4

Fig. 2. Finding agents that inhibit some of the genes that are upregulated on a particular Parkinson disease microarray.

The HSP27 (HSPB1) gene, which is over-expressed in the experimental results, has already been implicated in the pathogenesis of PD [22]. We identified paclitaxel and quercetin as substances that inhibit the expression of this gene. Paclitaxel has been identified and used as an antineoplastic agent due to its unique activity as a

microtubule-stabilizing agent. Interestingly, microtubules appear to be critical for the survival and death of nigral DA neurons, which are selectively affected in PD. Quercetin is a multipotent bioflavonoid with great potential for the prevention and treatment of disease. There is evidence of various *in vivo* and *in vitro* effects of quercetin, including anti-inflamatory, antioxidative, and potentially antineurodegenerative effects relevant to PD.

Paclitaxel	INHIBITS	HSPB1|HSPB1 protein, human
Paclitaxel inhibits expression of **heat shock protein 27** (PMID: 15304155)		
Paclitaxel (Pacl) was reported to **suppress HSP27** (PMID: 19080259)		

Quercetin	INHIBITS	HSPB1|HSPB1 gene
Quercetin …, **inhibited** the expression of both HSP70 and **HSP27** (PMID: 12926076)		

4.3.2 Stimulate the downregulated

Our approach also provides interesting results when we search for substances that stimulate downregulated genes in the transcriptomic experiment. For example, it turns out that *Pramipexol* stimulates expression of *NR4A2*. Pramipexol is a non-ergotic D2/D3 dopaminergic agonist that can be used to treat the symptoms of PD safely and effectively, both as monotherapy in the early stages and in the advanced phases in association with levodopa. Furthermore, in laboratory studies pramipexole exerts neuroprotective effects and its use has been related to a delay in the appearance of motor complications.

NR4A2 (Nurr1) encodes a member of the steroid-thyroid hormone-retinoid receptor superfamily. The encoded protein may act as a transcription factor. Mutations in this gene have been associated with disorders related to dopaminergic dysfunction, including PD. Nurr1 has been shown to be involved in the regulation of alpha-synuclein. Decreased expression of Nurr1, which has been found in PD patients with Nurr1 mutations, increases alpha-synuclein expression.

pramipexol	STIMULATES	NR4A2
… the increase of **Nurr1 gene** expression **induced by PRX**, ... (PMID: 15740846)		
… the **induction of Nurr1 gene** expression by **PRX** ... (PMID: 15740846)		

NR4A2	ASSOCIATED_WITH	Parkinson Disease
… lower levels of **NURR1** gene expression were **associated** with significantly increased risk for **PD** (PMID: 18684475)		

We also found that *leptin* stimulates *CDC42*. This gene, which is downregulated in the transcriptomic experiment, codes for a protein which is a small GTPase. Recent data indicate that components of small GTPase signal transduction pathways may be

directly targeted by alpha-synuclein oligomers, which potentially leads to signaling deficits and neurodegeneration in PD. Leptin on the other hand is a hormone secreted from white adipocyts. There is evidence that leptin prevents the degeneration of dopaminergic neurons by 6-OHDA and may be useful in treating PD.

5. Conclusion

In this paper we presented an application that integrates the results of microarray experiments with a large database of semantic predications representing the content of nearly 5 million MEDLINE citations. We discuss the value of this system with examples from microarray data on Parkinson disease, illustrating the way semantic relations elucidate the relationship between current knowledge and information gleaned from the experiment and help generate novel hypotheses.

References

1. Cordero F, Botta, M, Calogero RA.: Microarray data analysis and mining approaches. Brief Funct Genomic Proteomic.6, 265--281 (2007)
2. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 36, 462--77 (2003)
3. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R.: NCBI GEO: Mining tens of millions of expression profiles - database and tools update. Nucleic Acids Res.35 (Database issue), D760--D765 (2007)
4. Shatkay H, Edwards S, Wilbur WJ, Boguski M.: Genes, themes and microarrays: using information retrieval for large-scale gene analysis. Proc Int Conf Intell Syst Mol Biol. Pp. 317--28. (2000)
5. Blaschke C, Oliveros JC, Valencia A.: Mining functional information associated with expression arrays. Funct Integr Genomics. 1, 256--268 (2001)
6. Yang J, Cohen AM, Hersh W.: Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts. In: AMIA Annu Symp Proc., pp. 831--835 (2007)
7. Leach SM, Tipney H, Feng W, Baumgartner WA, Kasliwal P, Schuyler RP, Williams T, Spritz RA, Hunter L.: Biomedical discovery acceleration, with applications to craniofacial development. PLoS Comput Biol. 3, e1000215 (2009)
8. Jelier R, 't Hoen PA, Sterrenburg E, den Dunnen JT, van Ommen GJ, Kors JA, Mons B.: Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. BMC Bioinformatics. 9, 291 (2008)

9. Burkart MF, Wren JD, Herschkowitz JI, Perou CM, Garner HR.: Clustering microarray-derived gene lists through implicit literature relationships. Bioinformatics. 23, 1995--2003 (2007)

10. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 30, 7--18 (1986)

11. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM.: Using literature-based discovery to identify disease candidate genes. Int J Med Inform. 74, 289--98 (2005)

12. Hristovski D, Friedman C, Rindflesch TC, Peterlin B.: Exploiting semantic relations for literature-based discovery. In: AMIA Annu Symp Proc., pp. 349--353 (2006)

13. Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC.: Using the literature-based discovery paradigm to investigate drug mechanisms. In: AMIA Annu Symp Proc., pp. 6--10 (2007)

14. Masseroli M, Kilicoglu H, Lang FM, Rindflesch TC.: Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. BMC Bioinformatics. 7, 291 (2006)

15. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC.: Extracting semantic predications from Medline citations for pharmacogenomics. In: Pac Symp Biocomput., pp. 209--20 (2007)

16. Aronson AR.: Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In: Proc AMIA Symp., pp. 17--21 (2001)

17. Tanabe L, Wilbur WJ.: Tagging gene and protein names in biomedical text. Bioinformatics.18, 1124--1132 (2002)

18. R Development Core Team.: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2008)

19. Gentleman RC, et al.: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004)

20. Benjamini Y, Hochberg Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B.57, 289--300 (1995)

21. Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RK, Graeber MB.: Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics. 7, 1--11 (2006)

22. White LR, Toft M, Kvam SN, Farrer MJ, Aasly JO.: MAPK-pathway activity, Lrrk2 G2019S, and Parkinson's disease. J Neurosci Res. 85, 1288--1294 (2007)