Pacific Symposium on Biocomputing 11:367-378(2006)

MINING ALZHEIMER DISEASE RELEVANT PROTEINS FROM INTEGRATED PROTEIN INTERACTOME DATA

JAKE YUE CHEN[†]

Indiana University School of Informatics
Purdue University School of Science, Dept. of Computer and Information Science
Indianapolis, IN 46202, USA

CHANGYU SHEN

Division of Biostatistics, Indiana University School of Medicine
Indianapolis, IN 46202, USA

ANDREY Y. SIVACHENKO

Ariadne Genomics, Inc., 9700 Great Seneca Hwy
Rockville, MD 20850, USA

Huge unrealized post-genome opportunities remain in the understanding of detailed molecular mechanisms for Alzheimer Disease (AD). In this work, we developed a computational method to rank-order AD-related proteins, based on an initial list of AD-related genes and public human protein interaction data. In this method, we first collected an initial seed list of 65 AD-related genes from the OMIM database and mapped them to 70 AD seed proteins. We then expanded the seed proteins to an enriched AD set of 765 proteins using protein interactions from the Online Predicated Human Interaction Database (OPHID). We showed that the expanded AD-related proteins form a highly connected and statistically significant protein interaction sub-network. We further analyzed the sub-network to develop an algorithm, which can be used to automatically score and rank-order each protein for its biological relevance to AD pathways(s). Our results show that functionally relevant AD proteins were consistently ranked at the top: among the top 20 of 765 expanded AD proteins, 19 proteins are confirmed to belong to the original 70 AD seed protein set. Our method represents a novel use of protein interaction network data for Alzheimer disease studies and may be generalized for other disease areas in the future.

1. Introduction

Alzheimer Disease (AD) is a progressive neurodegenerative disease with 4.5 million patients in the United States today. This number of AD patients is

[†] To whom correspondence should be sent. Email: jakechen@iupui.edu.

expected to increase to 11 to 16 million by 2050 when the baby boomers age. The cognitive function of an AD patient deteriorates irreversibly over time and complete care is required for basic daily activities in the late stages of the disease. In 2000, health care costs for AD patients in the United States totaled approximately $31.9 billion, which is expected to reach $49.3 billion by 2010 (above statistics can be found at http://www.alz.org/). Therefore, AD is a major and rapidly growing public health concern.

The exact molecular mechanisms leading to the clinical symptoms and neuropathological changes associated with AD remain unclear. Selective brain neuronal loss, extracellular amyloid (senile) plaques, and intracellular neurofibrillary tangles (NFT) of hyperphosphorylated *tau* protein are characteristically seen in the brains of AD patients [*1, 2*]. According to the widely-accepted "amyloid hypothesis" [*3, 4*], an unusual accumulation of beta-amyloid peptides (Aβ), cleavage products of the amyloid precursor proteins (APP), are the major cause of AD in its earliest stages. In *Familial Alzheimer Disease* (FAD), genetic defects code for abnormal variants of either the APP or presenilin (PSEN1, PSEN2)— often leading to abnormal formation of Aβ as "protofibrils" [*5*]. Aβ protofibrils can incite inflammatory response through cytotoxic cytokines and disrupt intracellular Ca^{2+} homeostasis through over-activation of glutamate receptors, therefore leading cells to oxidative stress and mitochondrial injury. Aβ protofibrils deposit in the extracellular space which may also cause neuronal cell damage by blocking axonal transport. Aberrant Aβ accumulation further causes aberrant accumulation of *tau*, a protein which normally is essential to the initiation and stabilization of the neuronal microtubules. As time going by, gradual breakdown of neuronal cytoskeleton eventually leads to neuron apoptosis in AD patients (For a comprehensive review, see [*1, 2*] and references therein). The complexity and broad range of these cellular and biochemical events make researchers believe that there must be a sophisticated network of AD signal transduction, gene regulation, and protein-protein interaction events. Therefore, deciphering AD-related molecular network "circuitry" can help researchers understand AD disease model details and propose treatment ideas.

In this work, we will conduct initial AD-protein interaction network analysis and demonstrate how to gain protein functional knowledge not directly implied from sequence information. We will organize the main body of the work by presenting our computational data analysis methods and results. We will discuss potential interpretations and significance of our results at the end.

2. Computational Method

We introduce the computational techniques and procedures developed for AD protein interaction sub-network analysis, which can be summarized as follows. First, we searched the Online Mendelian Inheritance in Man (OMIM) database [*6*] to obtain an initial collection of AD-related genes. Second, we used the HUGO Gene Nomenclature Committee (HGNC) [*7*] database to map the initial AD-related genes to AD-related proteins identified by their SwissProt IDs. Third, we used a nearest-neighbor expansion method to build an expanded AD protein interaction sub-network. Fourth, we developed and applied a bioinformatics software tool, ProteoLens [*8*], to visualize and annotate the AD interaction sub-network. Fifth, we performed statistical analysis to assess the significance of the subnetwork extracted. Sixth and lastly, we developed a heuristic algorithm and scoring method, which we used to obtain a rank-ordered list of proteins significantly related to the AD. A detailed description of our method is provided below.

2.1. *Initial Collection of AD Related Genes*

We used the OMIM database [*6*] as the starting point to retrieve an initial collection of AD related genes. In OMIM, human genes associated with genetic disorders are recorded in a mini-review format, along with additional information such as their functions, participating molecular pathways, and other disease-related information. To obtain a list of AD-related genes, we performed a search of the OMIM database (integrated into our biological data warehouse in early 2004), retrieving each OMIM gene record in which the "description" field contains the term "Alzheimer". 65 OMIM gene records were retrieved. Note that since the retrieval method is coarse (i.e., based on simple term matches), the 65 collected AD-related gene records may suffer from both *false positives* (containing retrieved genes that are not actually functionally relevant to AD) and *false negatives* (missing genes that are indeed functionally related to AD but not retrieved). Soon in subsequent protein interaction network analysis, we will use protein interaction network neighborhood information and show that these concerns can be ameliorated.

2.2. *Mapping of Initial AD Related Genes to Proteins*

We used the HUGO Gene Nomenclature Committee (HGNC) [*7*] database of gene symbols and proteins to map gene symbols to their correct protein identifiers. HGNC is an international standard repository of officially approved gene symbols. For each gene, the HGNC database provides its standard gene symbol and gene mappings to various IDs used in common public databases,

e.g., Swiss-Prot, NCBI RefSeq, NCBI Locuslink, and KEGG enzyme. For our work, we started with 65 sets of OMIM gene records, some of which were associated with more than one gene symbol. After mapping all the gene symbols to protein SwissProt IDs using the HGNC gene mapping table, we obtained 70 AD-related proteins. The slight increase in protein count is due to one-to-many mapping between a gene and its multiple splice variant forms at the protein level.

2.3. *Collection and Expansion of AD Related Protein Interactions*

We used the Online Predicted Human Interaction Database (OPHID) [*9*] to collect AD-related protein interaction data. OPHID is a web database of more than 40,000 human protein interactions involving ~9,000 human proteins. It is a comprehensive repository of known human protein interactions, both from curated literature publications and from high-throughput experiments. It also contains predicted interactions inferred from eukaryotic model organisms, e.g., yeast, worm, fly, and mouse. The prediction was performed by mapping interacting protein pairs from available model organisms onto their orthologous protein pairs in human, or by making inference from interacting domain co-occurrence, co-expression, and GO semantic distance evidence. More than half of OPHID's records are predicted human protein interactions; however, not all OPHID human protein interactions carry the same level of significance. In general, those derived from real human protein interaction experiments should be much more trustworthy than those derived from predictive methods applied on yeast data sets. Therefore, to assign an estimated interaction confidence score, we developed the following *heuristic* scoring rules:

1. Protein interactions from human experimental measurement or from literature curation are assigned a **high** confidence score of 0.9;
2. Human protein interactions inferred from high-quality interactions in mammalian organisms are assigned a **medium** confidence score of 0.5;
3. Human protein interactions inferred from low quality interactions or non-mammalian organisms are assigned a **low** confidence score of 0.3.

With the initial AD-related protein list and a comprehensive OPHID protein interaction data set, we can now derive a AD-related protein interaction sub-network using a **nearest-neighbor expansion method**. Here, we denote the initial 70 AD-related proteins as the **seed-AD-set**. To build AD sub-networks, we pulled out protein interacting pairs in OPHID such that at least one member of the pair belongs to the seed-AD-set. The set of interacting pairs pulled out

will be called the **AD-interaction-set**. We denote the new set of proteins expanded from initial seed-AD-set by new proteins involved in the AD-interaction-set as the **enriched-AD-set** (a superset of seed-AD-set). In our study, the AD-interaction-set contains 775 human protein interactions; the enriched-AD-set contains 657 human proteins identified by Swissprot IDs.

2.4. *Visualization of AD Protein Interaction Sub-Network*

We developed ProteoLens [*10*], a visual biological network data mining and annotation tool that can be freely downloaded at http://bio.informatics.iupui.edu/proteolens/, to help us analyze the AD-related protein interaction sub-network. ProteoLens has native built-in support for relational database access and manipulations. It allows expert users to browse database schemas and tables, filter and join relational data using SQL queries, and customize any combination of data fields. The reconfigured view of data can be immediately visualized in the ProteoLens network viewer without needing to be exported as flat files first. Note that network nodes and edges can be used to represent proteins and protein interactions, whereas node/edge size, width, shape, and color can all be used to dynamically bind to customized data fields (such as gene symbol, functional category, and confidence score) to be visualized. Once a visual network layout is generated, the layout, visual annotation, and network member proteins/protein interactions can be tweaked without file editing.

2.5. *Statistical Evaluation of Sub-network*

We performed statistical data analysis tests to examine the significance of the connected sub-network formed by AD-interaction-set. Our hypothesis for this statistical evaluation is that if the enriched-AD-set indeed consists of functionally related proteins involved in the same process—even if the process were complex and broad—then we should expect that the **connectivity** among the enriched-AD-set proteins to be higher than that among a set of randomly selected proteins.

To formulate our hypothesis precisely, we introduce three concepts. First, we define a **path** between two proteins A and B as a set of proteins P1, P2,…, Pn such that A interacts with P1, P1 interacts with P2, …, and Pn interacts with B. Note that if A directly interacts with B, then the path is the empty set. Second, we define the **largest connected sub-network** of a network as the largest subset of proteins and interactions, among which there is at least one path between any two proteins in the subset. Third, we define the **index of aggregation** of a network as the ratio of the size of the largest sub-network that

exists in this network to the size of this network. Note that size is calculated as the total number of proteins within a given network/sub-network.

To test the hypothesis that the enriched-AD-set proteins are "more connected" than a randomly selected set of protein, we develop the null hypothesis test using the following resampling procedure [*11*]:

1) Randomly select from the OPHID database, the same number of human proteins as in the seed-AD-set.
2) Build the superset of the selected set by using the same nearest-neighbor expansion method described earlier.
3) Find the largest sub-network of the superset.
4) Compute the index of aggregation of the superset.
5) Repeat steps 1 through 4 1,000 times to generate a distribution of the index of aggregation under random selection.
6) Compare the index of aggregation of the enriched-AD-set with the distribution obtained in 5 and calculate the *p*-value.

2.6. *Scoring of Significant Proteins in the Sub-network*

In the final step, we present a scoring method to rank proteins in the sub-network, based on their overall roles and contribution to the AD related protein interaction sub-network. The **role** of a protein in the sub-network can be qualitatively defined as its ability to connect to many protein partners in the network with high specificity (the less promiscuously connected, the better) and high fidelity (the higher the interaction confidence, the better). To define this role quantitatively, we introduce a heuristic **relevance score** function s_i for each protein i from the sub-network:

$$s_i = k * \ln\left(\sum_{j \in N(i) \cap A} p(i,j)\right) - \ln\left(\sum_{j \in N(i) \cap A} N(i,j)\right), \quad (1)$$

In Eq. 1, i and j are indices for proteins in the sub-network, k is an empirical constant (k>1; we set k=2 here), $N(i)$ is the set of interaction partners of protein i in the network, A is the set of proteins in enriched-AD-set, $p(i,j)$ is the initial confidence score that we assigned to each interaction between proteins i and j (described in section 2.3), and $N(i,j)$ holds the value of 1 if protein j belongs to the intersection of $N(i) \cap A$ or 0 otherwise. Empirically assessing the relevance score function, we can tell that the score s_i ranks favorably in situations where interacting proteins with many high confidence interactions among its neighbors will fare out better than those with many low-quality interactions and those with only a few interactions. To avoid showing a negative score, in this work, we further converted s_i to the exponential scale using the transformation $t_i = exp(s_i)$, and report t_i as the final protein ranking score.

3. Results

By following the data analysis steps outlined in the Method section, we obtained the following results.

3.1. *AD-Related Proteins and Protein Interactions*

In the AD seed set, we have an initial list of 65 AD related OMIM gene records. These records are later mapped to 70 seed-AD-set proteins using gene-to-protein mapping tables from HGNC. As explained earlier, this discrepancy was due to the one-to-many mapping relationships between genes and their protein products. Using OPHID and the nearest-neighbor expansion method, we obtained 775 AD-related human protein interactions (as the **AD-interaction-set**). This expanded AD-interaction-set contains an expanded 657 human proteins (as the **enriched-AD-set**).

The proteins in the enriched-AD-set form 16 sub-networks, with a size ranging from 2 to 586 (or, a relative size from 0.3% to 89.2%). Therefore, the largest connected sub-network of the enriched-AD-set contains 586 proteins and the **index of aggregation** is 82.9%. This suggests that the majority of AD related proteins are closely related by physical interaction—a phenomena that we would like to test for statistical significance (see Section 3.3)

3.2. *Visualization of AD Expanded Protein Interaction Network*

Figure 1 shows the enriched-AD-set proteins and the AD-interaction-set of human protein interactions in visualized network. All the seed-AD-set proteins (shown as nodes) are colored dark gray, while the non seed-AD-set proteins (also shown as nodes) are colored light gray. Proteins with high ranking scores (see Table 1 and the Discussion section) are also draw as nodes with sizes proportion to the ranking score. All the protein interactions (shown as edges) are also color-labeled, with high-quality interaction in black, medium-quality interactions in dark gray, and low-quality interactions in gray. We observe that interactions tend to "fan out" from a few protein hubs in the network, and that there are relatively few interactions among the proteins extending from the seed-AD-set. One expects that true AD-related proteins would interact with many seed-AD-related proteins with high degree of confidence and specificity.

3.3. *Statistical Significance*

The empirical distribution of the index of aggregation obtained after 1000 random re-samplings is shown in Figure 2. Only 8 runs out of 1000 resulted in an index of aggregation value greater than 89.2%. Therefore, the *p*-value of the

observed index of aggregation of the enriched-AD-set is 0.008. It is not surprising to observe such a significant result since the AD-set is selected in a way that proteins inside the set share certain level of connection since all of them are associated with AD.

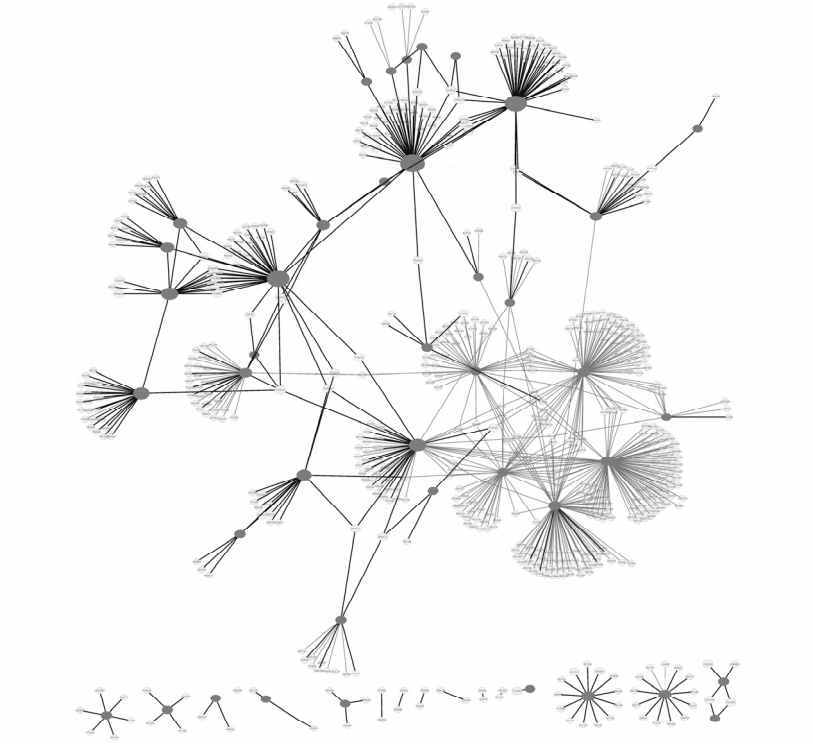

Figure 1. A network of OPHID human interactions expanded from initial 70 seed-AD-set (colored in dark gray). Protein interactions are colored in different shades of gray according to confidence level assigned, and protein node size as shown in proportion to their significance in the network (see text for details).

Next, the relevance score was calculated for each protein in the enriched-AD-set. The results (Table 1) show that our scoring function exhibits very high specificity: out of 20 top-scoring proteins, all but one (β-catenin, CTNNB1) are known AD-related proteins according to OMIMM annotation. Further literature study (see discussion) suggests that even CTNNB1 could be involved in the AD disease development process [*12*]. This result opens up exciting new possibilities to identify novel or previously ignored members of the AD pathway(s) for subsequent protein drug target investigations or disease biomarker studies.

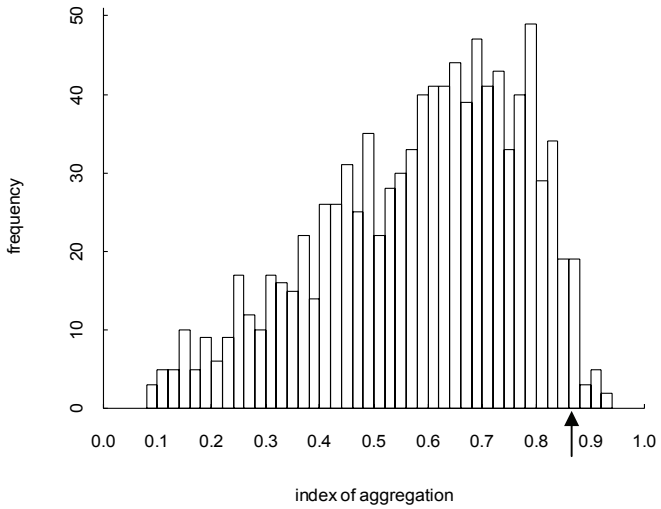

Figure 2. Histogram of the index of aggregation distribution for the enrichments of sets of proteins (size=48) randomly selected from OPHID. The arrow indicates the index of aggregation value for the enriched-AD-set.

4. Discussion

The integrated approach to the analysis of human interaction data and Alzheimer disease proteins allowed us to validate existing disease protein targets and predict novel ones not present in the initial list of disease protein targets that we started with. The result can be interesting to Alzheimer disease biologists, and our method can be generalized to other disease biology areas. By further examining our discoveries in the top-ranked proteins (top 20 are tabulated in Table 1, other will be made available on our web site http://bio.informatics.iupui.edu/ soon), we can make the following comments:

First, one of the important Alzheimer disease proteins, tau protein (MAPT, ranked 31), a well-known participant in AD-linked degeneration pathway(s), was not initially retrieved by our automated procedure from OMIM data but was later recovered from the interaction data analysis. Therefore, at least in a few isolated cases, our method can allow the recovery of false negatives.

Second, the amyloid beta A4 precursor-protein binding protein, APPB1 (ranked 33), represents another interesting case. It is a well-known interaction partner of APP, but a genetic link to AD was reported in OMIM only for the other member of the family, APPB2 (ranked 32). Our method still predicts that APPB1 also plays some role in AD. A recent literature report [*13*], shows that

APPB1 indeed directly associates with tau and may provide the crucial missing link between tau and APP proteins in Alzheimer disease.

Table 1. Top 20 rank-ordered AD relevant proteins.

Score	Gene	Description	AD Relevance
43.01	APP	amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)	Known
36.98	PSEN1	presenilin 1 (Alzheimer disease 3)	Known
35.64	LRP1	low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor)	Known
21.87	PSEN2	presenilin 2 (Alzheimer disease 4)	Known
20.89	PIN1	protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting 1	Known
19.37	FHL2	four and a half LIM domains 2	Known
15.39	S100B	S100 calcium binding protein, beta (neural)	Known
12.96	FLNB	filamin B, beta (actin binding protein 278)	Known
12.37	CTNND2	catenin (cadherin-associated protein), delta 2 (neural plakophilin-related arm-repeat protein)	Known
12.15	CLU	clusterin (complement lysis inhibitor, SP-40,40, sulfated glycoprotein 2, testosterone-repressed prostate message 2, apolipoprotein J)	Known
11.34	APBA1	amyloid beta (A4) precursor protein-binding, family A, member 1 (X11)	Known
10.00	NAP1L1	nucleosome assembly protein 1-like 1	Known
9.54	GTPBP4	GTP binding protein 4	Known
9.48	NCOA6	nuclear receptor coactivator 6	Known
9.15	CDK5	cyclin-dependent kinase 5	Known
7.44	CTSB	cathepsin B	Known
7.29	ASL	argininosuccinate lyase	Known
4.86	**CTNNB1**	**catenin (cadherin-associated protein), beta 1, 88kDa**	**Novel**
4.86	NCKAP1	NCK-associated protein 1	Known
4.86	AGER	advanced glycosylation end product-specific receptor	Known

Third, β-Catenin (CTNNB1, ranked 18 as shown in Table 1), was not previously associated with AD in OMIM or in the general biomedical community, therefore representing a "clear" case of computational prediction results. Interestingly, while the exact role of β-catenin in AD is not well understood, it is known that Wnt signaling pathway (which β-catenin is a part of) is a target of Aβ toxicity [*14*]. Moreover, the Wnt-3a ligands and other agents that are reported to overcome beta amyloid toxicity stabilize CTNNB1 levels in cytoplasm [*15, 16*]. It should be stressed that while the OPHID interactions between β-catenin and AD-set proteins are of high-quality, *i.e.* derived from the literature, one could only speculate about the potential role of β-catenin, since, for instance, both CTNNB1 and its interaction partners's

expressions are far from being limited to neurons. It is the *pattern* of β-catenin interactions revealed through the analysis of combined evidence that resulted in high AD-relevance score for β-catenin.

In all, our method incorporated protein interaction data and helped us to successfully carry out Alzheimer disease-related biological studies. The computational results, which began with inputs that are not necessarily highly reliable, showed high biological relevance. Going down the ranked protein targets, one may generate many new biological hypotheses about the new functions of proteins in the protein interaction network context beyond the scope of this work. We are currently developing collaborations with industrial partners who have accumulated experimental human protein interactions? to further conduct these computational investigations on high-confidence data sets. Meanwhile, we are also in the process of developing better scoring functions and applying these methods to the study of other disease areas.

Acknowledgments

This work was supported in part by systems obtained by Indiana University through its relationship with Sun Microsystems Inc. as a Sun Center of Excellence. We would like to thank Stephanie Burks who helps us maintaining a robust computer systems and Oracle 10g database server at Indiana University.

References

1. E. Bossy-Wetzel, R. Schwarzenbacher and S. A. Lipton, *Nat Med* **10 Suppl**, S2 (2004).
2. A. Kowalska, *Pol J Pharmacol* **56**, 171 (2004).
3. J. Hardy, *Trends Neurosci* **20**, 154 (1997).
4. D. J. Selkoe, *Nature* **399**, A23 (1999).
5. M. Hutton, J. Perez-Tur and J. Hardy, *Essays Biochem* **33**, 117 (1998).
6. . McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD); Web URL: http://www.ncbi.nlm.nih.gov/omim/ 2000.
7. S. Povey, et al., *Hum Genet* **109**, 678 (2001).
8. A. Sivachenko and J. Y. Chen, *Bioinformatics* (2005), submitted.
9. K. R. Brown and I. Jurisica, *Bioinformatics* **21**, 2076 (2005).
10. J. Y. Chen and A. Sivachenko, *IEEE Magazine in Biology and Medicine* **24**, 95 (2005).
11. W. J. Ewens and G. Grant, "Statistical Methods in Bioinformatics: An Introduction" Springer, New York, 2004.

12. M. Nishimura, et al., *Nat Med* **5**, 164 (1999).
13. C. Barbato, et al., *Neurobiol Dis* **18**, 399 (2005).
14. R. A. Fuentealba, et al., *Brain Res Brain Res Rev* **47**, 275 (2004).
15. A. R. Alvarez, et al., *Exp Cell Res* **297**, 186 (2004).
16. R. A. Quintanilla, et al., *J Biol Chem* **280**, 11615 (2005).