The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
Copyright © 2008 ORSC & APORC, pp. 232–241

Analysis of domain interactions among gamma-secretase substrates

Zhi-Ping Liu[1,*] Weiming Xia[2,†] Luonan Chen[3,4,‡]

[1]Institute of Applied Mathematics, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
[2]Center for Neurologic Diseases, Brigham and Women's Hospital,
Harvard Medical School, Boston, Massachusetts 02115, USA
[3]Department of Electronics, Information and Communication Engineering
Osaka Sangyo University, Osaka 574-8530, Japan
[4]Institute of Systems Biology, Shanghai University, Shanghai 200444, China

Abstract The γ-secretase plays a key role in the Amyloid hypothesis of the cause of Alzheimer's disease. The integral membrane protein cleaves single-pass transmembrane proteins at residues within the transmembrane domain. In this work, we proposed a new model to analyze the relationships among the identified substrates of γ-secretase at the domain level. Firstly the sequence features in domains of γ-secretase and its substrates are identified respectively. Based on the correlation analysis, we classified the domains of the related substrates into several groups according to their similarities. Then, we revealed the relationships between γ-secretase and its substrates in an interaction model and identified the sequence signatures of the domain clusters. What we found can be regarded as the interaction markers with γ-secretase. The domain features with the sequence signatures can be used not only to decipher the interaction mechanisms between the γ-secretase and their corresponding substrates but also to predict new protein-protein interactions related to γ-secretase.

Keywords γ-secretase; protein interaction; domain interaction; systems biology; Alzheimer's disease.

1 Introduction

Alzheimer's disease (AD) is a debilitating neurodegenerative disorder affecting about 24 millions of elderly individuals worldwide today [3]. The genetics mechanism causing the disease is still unclear and there are mainly three hypotheses about it [7, 17]. The most early one is about cholinergic hypothesis. It regarded AD as a result of reduced biosynthesis of neurotransmitter acetylcholine [17]. A new hypothesis of the mechanism is about the Tau proteins. Hyperphosphorylated Tau begins to pair with other threads of Tau and then they become tangled up together inside neuron cell in masses known as neurofibrillary tangles (NFT). When this happens, the microtubules disintegrate, collapsing the

[*]Email: zpliu@amss.ac.cn
[†]Corresponding author. Email: wxia@rics.bwh.harvard.edu
[‡]Corresponding author. Email: chen@eic.osaka-sandai.ac.jp

neuron's transport system [11]. The most popular one is the amyloid hypothesis. Mutations in the amyloid precursor protein (APP) and presenilin genes increase the production of peptide called Aβ42, which is the main component of senile plaques (SP) in brains of AD patients [14]. γ-secretase plays a fundamental role in AD by catalyzing the final proteolytic cleavage, which leads to the formation of amyloid-β peptide (Aβ) directly, the major component of the diseases defining SP [3]. In addition to the presenilins (PS1 and PS2), nicastrin (Nct), APH-1 and PEN-2 were recently identified to be subunits of the complex [9, 4]. Apparently, all four proteins assemble into a large 500-600-kDa complex, which displays the intramembranous proteolytic activity required for the cleavage of the β-amyloid precursor protein (βAPP) and other substrates such as Notch [7]. γ-secretase is a very versatile protease which cleaves many type I transmembrane proteins. More than 50 substrates that have been reported are cleaved by γ-secretase [3]. The interaction between γ-secretase and substrate proteins is involved in positioning the substrate into the initial docking site, which is spatially distinct from the catalytic site [3]. The specificities of interaction sites play a key role in modulating the substrates. Cleavage of these substrates yields intracellular domains that translocate into nuclei and control the expression of downstream target genes.

Domain-domain interaction (DDI) is a decisive factor for observed protein-protein interaction (PPI) [5]. Recognizing the interaction patterns between a domain and its parters would gain biological insight into understanding mechanism of protein networks, developing drugs to inhibit pathological protein interactions and designing novel protein interactions from appropriate domain scaffolds [13]. Moreover, a protein often consists of multiple domains. Therefore, it is a difficult task to infer DDI from the known PPI [6, 13] although it will provide important direction to investigate the interface or binding features among these proteins. In AD study, this will provide details of functional pathway that cause the disease signal transduction and protein-protein interaction events [5]. Therefore, discovering the interacting manner between γ-secretase and its substrates and detecting the common features of functional domains can help researchers understand AD disease model in a detail manner and further propose new treatment ideas.Sequence signature can be regarded as a conserved region and a repeat sequence pattern lies in a group of related protein sequences [6]. Thus, the discovery of these smaller sequence signatures allows researchers to structurally characterize PPI with more precision, and the significant motifs would provide detailed mechanisms for γ-secretase binding its substrates.

In this work, we proposed a new model to systematically analyze γ-secretase related substrates from their domain relationship. We aim to decipher the mechanisms of PPI between γ-secretase and its substrates from DDI and sequence signature perspective. We initially collect the sequences of γ-secretase and its contacting proteins. The domain configurations in every protein are identified individually. To explore the relationships among the domains of these substrates, we group the domains into several clusters from their sequence similarity by a hierarchical clustering process. The module organization of the clusters reveals the similarity patterns among the domains related to γ-secretase. To investigate the signature motifs in γ-secretase related domains, we identify the significant sequence signatures of every domain clusters. The identified sequence signatures of domains can be used to predict new protein-protein interactions between γ-secretase and its new substrates. In this paper, we show that the system level analysis will provide new insight into investigating AD mechanism which may directly be applied to drug design and other thereputical applications.

2 Results

2.1 Domain features

Firstly, we summarized the domains in the four subcomponents of γ-secretase. We identified four domains in the three of the four subunits from Pfam database [2]. Table 1 lists the domains of gamma-secretase in details, where there are no significant domains detected for PEN-2 till now. We mainly find sequence patterns among the domains of γ-secretase by multiple sequence alignment (MSA) [8]. From Figure 1 of the alignment result, clearly there are 3 mainly aligned regions among four domains of γ-secretase. The first one locates from the residues 15 to 29, the second is from 48 to 227 and the third is from 317 to 370. The similarity among the domains indicates the detailed analysis of the interfaces with binding substrates as well as the similar binding sites among the four components.

Table 1: Domains of gamma-secretase with high significance.

Component		Pfam ID	Pfam accession	Range/sequence length
PS	PS-1	Presenilin	PF01080.8	70-458/467
	PS-2	Presenilin	PF01080.8	76-439/448
Nct	–	Nicastrin	PF05450.6	274-499/709
Aph-1	–	Aph-1	PF06105.3	1-238/238
Pen-2	–	–	–	–/101

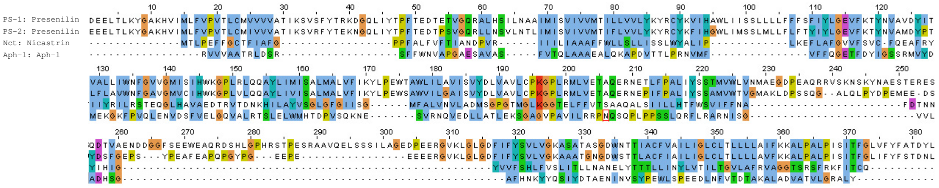


Figure 1: MSA among domains of γ-secretase.

As to the substrate proteins of γ-secretase, we identified their significant domains individually by a P-value threshold (see Methods). Some of them contained several significant domains. The domain length varies from 22 to 412 and the detailed list is available upon request from the authors. By comparing the residues of domain part with those of the corresponding protein, we can identify vital sequence features of domain within total sequence. Figure 2 shows the comparison results. We found that the fraction of hydrophobic residues (A, V, I, L, F) are favored in the domains. We also statistically compared the residue composition percentage in the substrates and that of their domains. The results show that there are no distinctive differences between them. When compared with that of γ-secretase, the hydrophobic residues have higher propensity in γ-secretase. It is interesting to find that the composition percentage of Leucine (L) in γ-secretase is obviously higher than that in the substrates, as well as the percentage of Cysteine (C) just in the opposite situation. Actually, it has reported that the Cysteine is important for local structures stabilizing conformation and performing functions [12].

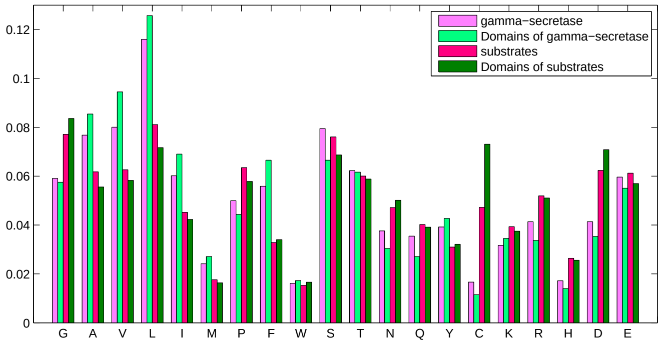


Figure 2: The composition percentage of amino acid residues in γ-secretase, its domains, its substrates and domains of the substrates.

2.2 Sequence clustering

To explore the relationships among the substrates of γ-secretase, we proposed a scheme to clustering the domains in the substrates. The domain correlations bridge the gaps of relationship among these proteins. We first built the distance matrix of the domains by their sequence dissimilarity which are measured by Smith-Waterman algorithm [15]. Then we grouped the identified domains by a hierarchical clustering with dynamic cutting of the dendrogram (see Methods). The dendrogram with a color map for every clustered modules is shown in Figure 3. The grey color modules show that these branches cannot be grouped into any modules significantly and should be regarded as isolated parts [10, 18]. There are totally 9 obvious clusters (9 colors in the color map) and all the rests are merged into a special grey module themselves.

Table 2 records the details of the identified 9 modules. We annotated every clusters with the domain IDs with assessment of reliability by the hypergeometric test (see Methods). The P-value is the probability that a cluster would be enriched of domains with a particular domain ID by chance alone. Smaller P-value indicates that cluster is significantly enriched for the specific domain ID and can be considered to be a module of the domain with high possibility.

The domains of substrates are clustered into similar groups and the relationships between the substrates can then be investigated from these clusters. We mapped the 55 type I membrane proteins to the clusters by annotating the frequency of hitting the domains contained in every clusters. The return mapping is shown in Figure 4, where we can find that the mapping profiles of homologue proteins are similar. For example, the homolog Notch1, Notch2 and Notch3 present similar profiles. It is known that they have similar functional performance during cooperation with γ-secretase [3]. The result provides an evidence that the analysis of domain correlations is effectiveness to examine the relation-

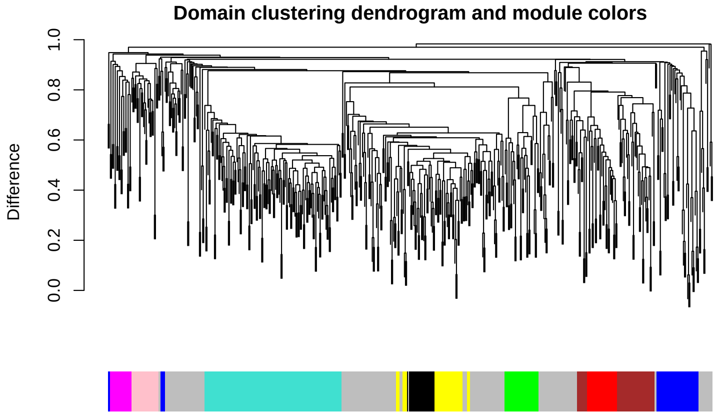


Figure 3: The dendrogram of clustering the domains in γ-secretase substrates and the color map of the identified modules after partitioning the hierarchy tree dynamically.

Table 2: The identified clusters with the significant domains in the cluster.

Cluster ID	Size	Domain ID	Frequency	P-value
1	121	PF00041.12	26	4.4e-016
		PF07679.7	10	6.2e-006
		PF00057.9	45	7.4e-005
2	72	PF00008.18	68	1.1e-016
3	55	PF07645.6	12	1.1e-005
4	53	PF00023.21	21	1.0e-032
		PF07714.8	6	3.4e-005
		PF00102.18	6	3.4e-005
5	49	PF00057.9	49	2.2e-016
6	39	PF00058.8	39	1.1e-016
7	37	PF00008.18	25	4.1e-007
		PF07645.6	11	1.6e-006
8	34	PF00058.8	20	6.9e-009
		PF01030.15	6	3.4e-006
9	21	PF00058.8	21	3.3e-016

ships among substrates, and also verifies our method. From the color mapping, we can easily grasp the scenario of associations among these substrates. In the figure, the module 0 (M0) represents the combined grey cluster in which the domains cannot be distinctly merged into any modules.

2.3 Sequence signatures of interaction

Proteins always perform their functions with interactions with other genetic components. Behind PPIs, there are protein domains interacting physically with one another to perform various functions [13], that is the same in AD related proteins [5]. PPI data

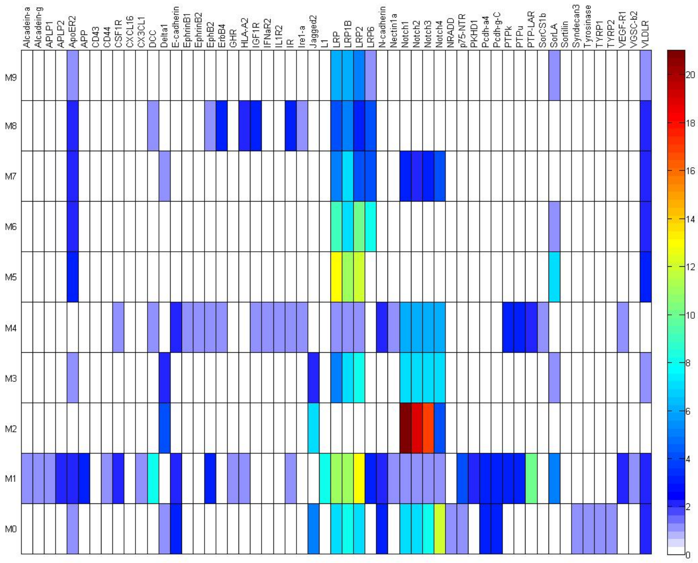


Figure 4: Mapping the substrates of γ-secretase to the clustered domain groups with the pseudocolor of frequency.

has been used to analyze DDI based on the widely accepted hypothesis that proteins interact with one another via conserved domains [13]. Large-scale PPI databases are used to identify correlated domains that are implicated in the binding of protein partners [13]. However, the binding sites of proteins are often much shorter than a domain, especially in the protein with long domain configurations [16]. To investigate the PPI at more detailed level for deciphering the mechanism of how the γ-secretase cleaves its substrates, we detected the sequence signatures underlying these substrates contacting with γ-secretase.

We regard sequence signatures as one of the decisive factors for protein interactions. In this case, there are so many "guest" proteins, i.e. the substrates, share a common interacting parter, i.e. the "host" γ-secretase. It will be crucial for discovering the mechanism of cleaving the substrates that to explore the sequence motifs as the interaction markers for binding with γ-secretase. Moreover, we have grouped the domains of substrates into similar clusters. Thus, we can detect the sequence signatures in every domain groups respectively. Here, sequence signatures shared by domain modules of guest proteins were initially discovered using the program MEME based on expectation maximization (EM) algorithm [1]. With regard to the detected significant sequence motifs of every cluster, Table 3 shows an example. The expectation values of every motif are listed in "E-value" column to assess its statistical significance. "Width" gives sequence length of these identified signatures. "Sites" records the number of domains containing target motif in the same cluster. The regular expression of these identified sequence signatures are also listed.

For these identified sequence signatures of substrates cleaved by γ-secretase, the conserved motifs will play essential roles in interacting with the same parter. The similar short sequence patterns among γ-secretase substrates imply the interaction markers. The

Table 3: The identified sequence signatures of domains underlying gamma-secretase related substrates.

Motif ID	Width	Sites	E-value	Motif regular expression
1	21	45	1.4e-349	RC[IV]PxS[KW][RVL]C[DN]GxDDCGDGSDE
2	29	23	4.5e-119	TL][ST]YT[ILV]xGLKP[DG]T[TE]YS[FI]RV[LQR]Ax[NT]xKGPGPP
3	37	10	1.9e-046	[PL][YR][FL][LT][KNQ]xP[SQE][SDN][HLQV][TV][AV]L[EP][GS][GS][TSD][AV]T[LF] [EDL]CQ[AV]SG[AE]P[MV]P[TES][IV][TK]W[LMF]K[NG]

rule of interactions with γ-secretase and the binding sites would lie in the position of these sequence motifs. When one of these sequence signatures is observed in a target protein, it is possible to predict its interaction with γ-secretase based on the knowledge of correlated domains and the sequence signatures. The host protein cleaves a number of proteins and product peptides to destroy the neuron systems. PPI with γ-secretase which would cause AD can be seized with these identified sequence signatures from the sequence perspective, especially when number of the transmembrane protein's structures still is not be available.

3 Discussion and Conclusion

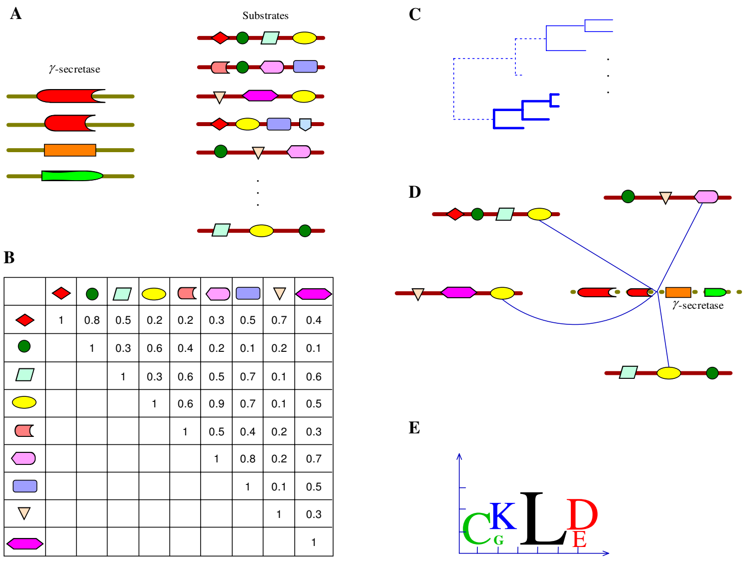
In this work, we gave a detailed correlation analysis of domains contained in γ-secretase substrates. The fundamental features of these domains were investigated and analyzed. The similarity among the domains of γ-secretase-related proteins were then used to group them into similar clusters. The identified clusters provide a foothold for further detection of the sequence signatures of the proteins interacting with γ-secretase. We identified the probabilistic sequence signatures as the interaction markers with γ-secretase. From such an analysis at system-wide level, we identified the features of domains and interaction markers. The sequence patterns in substrates of γ-secretase provided valuable insights into deciphering the mechanism of how γ-secretase cleaves its related proteins. The identified sequence features among the substrates of γ-secretase are the fundamental features with great potential for further intriguing research. The identified sequence features among the substrates of γ-secretase are the fundamental features with great potential for further intriguing research. We will integrate more domain information for these substrates and investigate the comprehensive interactions between the two terminals of host proteins and its guests as well as within them. The validation *in vivo* for these motifs is in preparation and will be proposed in another paper.

In summary, the paper proposed a novel framework to analyze functionally important domains and sequence motifs for relating γ-secretase. The proposed approach can be also viewed as an alternative to decipher the interaction mechanism among protein complexes of disease from the domain and sequence motif perspective. The results provide biological insight into deciphering the cleaving process of γ-secretase and further understanding pathophysiological mechanism of Alzheimer's disease.

4 Methods

4.1 Data sources

We use the 55 type I membrane protein substrates listed in the reference [3]. Together with the four identified subunits of γ-secretase, we retrieved their sequences from NCBI website (http://www.ncbi.nlm.nih.gov/) directly. Figure 5 is an overview of the analysis proposed in this paper. The domain configurations in the γ-secretase and its substrate proteins are identified by searching Pfam domain database, which detects the profiles of domain by HMM method [2]. We used Pfam release 22.0 version and filtered out significant domains by P-value threshold 10^{-4}. Figure 5 (A) gives an illustration of the domains in γ-secretase and its related substrates.



	◆	●	▱	⬭	◖	⬡	▭	▽	⬣
◆	1	0.8	0.5	0.2	0.2	0.3	0.5	0.7	0.4
●		1	0.3	0.6	0.4	0.2	0.1	0.2	0.1
▱			1	0.3	0.6	0.5	0.7	0.1	0.6
⬭				1	0.6	0.9	0.7	0.1	0.5
◖					1	0.5	0.4	0.2	0.3
⬡						1	0.8	0.2	0.7
▭							1	0.1	0.5
▽								1	0.3
⬣									1

Figure 5: A schematic representation of the interaction analysis among the γ-secretase and its substrates.

4.2 Sequence clustering

The identified domains are compared by Smith-Waterman algorithm [15] in all-against-all manner to detect the similarity among them. A hierarchical clustering process is followed to group the domains into clusters based on the similarity matrix of sequence identity normalized Z-score. Figure 5 (B) shows an example of the distance matrix and (C) gives a sketch of the constructed hierarchy tree. We use the Dynamic Tree Cut algorithm [10] to partition the dendrogram and build the domain modules. The cutting procedure lies in the spirit of a dynamic module detecting and analyzing the shapes of the branches of the dendrogram. The method can identify branches that could not be revealed using the static cut scheme [10]. The successful applications of the algorithm to detect complicated modules can be found in references [18]. These clustered domain groups are annotated with domain IDs by the frequency of the domain types occurring in the modules. The domain IDs which would be significantly enriched in these domains of a clusters are evaluated using a P-value based on the hypergeometric test. The P-value is defined as

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{d}{i}\binom{n-d}{m-i}}{\binom{n}{m}},$$

where n is the number of domains, d is the number of a particular domain ID, m is the module size and k is the number of domain ID in this module.

4.3 Detecting interaction markers

The functions of γ-secretase cannot be separated from the interactions with its substrates. The specificity of the particular binding domains and smaller specific sequence motifs plays essential roles in the interaction relationships. We regard the sequence signatures among the domains as interaction markers for γ-secretase interacting with its related proteins. Detecting the markers will provide valuable insights for the mechanism of cleaving proteins. Figure 5 (D) illustrates an example to show the associations between γ-secretase and its substrate and related domains. To eliminate the bias of virous domain types, we use the program MEME [1] with the default parameters to detect the consensus sequence signals among the related domains in the same clusters that we have grouped. MEME is a motif finding algorithm based on expectation maximization [1, 13], which implements an unsupervised learning and produces probabilistic sequence signatures shared by target sequences. The statistical significance of the motif is assessed by an expectation value (E-value). Sequence signatures discovered by MEME are gapless sequences with the best width and the most frequency of occurrence based on statistical models [1, 6]. Figure 5 (E) shows a diagrammatic representation of interacting signatures of γ-secretase substrates.

Acknowledges

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 10631070. Part of the authors are also supported by the Grant No. 5039052006CB from the Ministry of Science and Technology of China and by NSFC-JSPS collaborative project No. 10711140116. The authors would like to thank Prof. Xiang-Sun Zhang for helpful discussions and suggestions.

References

[1] Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: 369–373

[2] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-D141

[3] Beel AJ, Sanders CR (2008) Substrate specificity of γ-secretase and other intramembrane proteases. Cell Mol Life Sci 65: 1311–1334

[4] Campbell WA, Yang H, Zetterberg H, Baulac S, Sears JA, Liu T, Wong ST, Zhong TP, Xia W (2006) Zebrafish lacking Alzheimer presenilin enhancer 2 (Pen-2) demonstrate excessive p53-dependent apoptosis and neuronal loss. J Neurochem 96: 1423–1440

[5] Chen JY, Shen C, Sivachenko AY (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pac Symp Biocomput 11:367–378

[6] Fang J, Hassl RJ, Dong Y, Lushington GH (2005) Discover protein sequence signatures from protein-protein interaction data. BMC Bioinformatics 6: 277

[7] Hardy J, Selkoe DJ (2002) The amyloid hypothesis of Alzheimer's Disease: progress and problems on the road to therapeutics. Science 297: 353–356

[8] Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

[9] Kimberly WT, LaVoie MJ, Ostaszewski BL, Ye W, Wolfe MS, Selkoe DJ (2003) Gamma-secretase is a membrane protein complex comprised of presenilin, nicastrin, Aph-1, and Pen-2. Proc Natl Acad Sci USA 100: 6382–6387

[10] Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24: 719–720

[11] Lee HG, Perry G, Moreira PI, Garrett MR, Liu Q, Zhu X, Takeda A, Nunomura A, Smith MA (2005) Tau phosphorylation in Alzheimer's disease: pathogen or protector? Trends Mol Med 11: 164–169

[12] Liu ZP, Wu LY, Wang Y, Chen L, Zhang XS (2007) Predicting Gene Ontology functions from protein's regional surface structures. BMC Bioinformatics 8: 475

[13] Riley R, Christopher Lee C, Chiara Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. Genome Biology 6: R89

[14] Selkoe DJ (1996) Amyloid β-Protein and the genetics of Alzheimer's disease. J Biol Chem 271: 18295–18298

[15] Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197

[16] Sprinzak R, Margalit H (2001) Coorelated sequence-signatures as markers of protein-protein interaction. J Mol Biol 311: 681–692

[17] Wolfe MS, Haass C (2001) The Role of presenilins in gamma-secretase activity. J Biol Chem 276: 5413–5416

[18] Zhang B, Horvath S (2005) A general framework for weighted gene coexpression network analysis. Stat Appl Genet Mol Biol 4: Article 17