A JAVA-BASED TOOL TO ANALYZE THE FUNCTIONAL PROTEIN SEQUENCES OF GENES CAUSING ALZHEIMER'S DISEASE

KIRAN KUMAR REDDI[1], T.K.RAMA KRISHNA RAO[2], G.NAGESWARA RAO[2]

[1]Department of Computer Science, Krishna University, Machilipatnam, India.
[2]Department of IT, AITAM, Tekkali, India.

E-mail: [1]kirankreddi@gmail.com, [2]ramakrishnatk@gmail.com

ABSTRACT

Alzheimer's disease is a progressive neurodegenerative disorder characterized by deposition of amyloid plaques composed of aggregated amyloid beta plaques, and neurofibrillary tangles composed of hyperphosphorylated tau that leads to synaptic defects resulting in neuritic dystrophy and neuronal death. Diseases like Alzheimer's are believed to result from the accumulation of mis-folded proteins. Most folded proteins have a hydrophobic core in which side chain packing stabilizes the folded state, and charged or polar side chains on the solvent-exposed surface where they interact with surrounding water molecules. In the present study, we extracted huge amounts of data from various biological databases available online. It is found that there are 74 genes that may cause Alzheimer's disease .We evaluated the role of 74proteins that are likely to be involved in Alzheimer's disease by employing multiple sequence alignment using ClustalW tool and constructed a Phylogenetic tree using functional protein sequences extracted from NCBI. The study also analyzed the physical properties of amino acids of functional proteins using the computational tool developed in java. These in silico study results are useful for new therapeutic interventions particularly in the field of biomarker discovery.

Keywords: *Alzheimer's Disease, Phylogenetic Tree, Amino Acids, Physical Properties, Java Tool.*

1. INTRODUCTION:

Traditional genetics and molecular biology have been directed toward understanding the role of a particular gene or protein in an important biological process. A gene is sequenced to predict its function or to manipulate its activity or expression. In contrast, the availability of genome sequences provides the sequences of all the genes of an organism so that important genes influencing metabolism, cellular differentiation and development, and disease processes in animals and plants, can be identified and the relevant genes manipulated (AARao et al, 2006). A major application of Bioinformatics is analysis of the full genomes of organisms that have been sequenced starting in the late 1990s.In the present study, we focused on the genes or proteins that are believed to have a major role in the pathogenesis of Alzheimer's disease using bioinformatics tools. Alzheimer's disease is a progressive neurodegenerative disorder characterized by amyloid plaques composed of aggregated amyloid beta plaques,neurofibrillary tangles (NFT) composed of hyperphosphorylated tau and synaptic defects resulting in neuritic dystrophy and neuronal death (Hutton M and McGowan E, 2004). The precise amino acid content, and the sequence of those amino acids, of a specific protein, is determined by the sequence of the bases in the gene that encodes that protein. The chemical properties of the amino acids of proteins determine the biological activity of the protein. Proteins not only catalyze all (or most) of the reactions in living cells, they control virtually all cellular process. In addition, proteins contain within their amino acid sequences the necessary information to determine how that protein will fold into a three dimensional structure, and the stability of the resulting structure. It is well known that amino acids in peptides and proteins show individually distinct preferences for secondary structural conformations. While each amino acid is associated with multiple physicochemical properties, most existing statistical analysis methods treat amino acids separately and independently in calculating the frequencies of

residues at each site of protein sequences. Such approaches remain at describing coarse-level tendencies of amino acid preferences for particular conformations (Gouchol Pok et al, 2008). Here we performed computational analysis of the amino acid sequences from Homo sapiens. We have developed software using java language that carries out the analysis of different physical properties of amino acids in proteins. Diseases like AD are believed to result from the accumulation of mis-folded proteins. Most folded proteins have a hydrophobic core in which side chain packing stabilizes the folded state, and charged or polar side chains on the solvent-exposed surface where they interact with surrounding water molecules. It is generally accepted that minimizing the number of hydrophobic sidechains exposed to water is the principal driving force behind the folding process, although a recent theory has been proposed which reassesses the contributions made by hydrogen bonding The process of folding *in vivo* often begins co-translationally, so that the N-terminus of the protein begins to fold while the C-terminal portion of the protein is still being synthesized by the ribosome. Cells express specialized proteins called chaperones whose function is to aid in the folding of other proteins.

2. MATERIALS AND METHODS:

2.1 SEQUENCE COLLECTION:

The genes causing Alzheimer's disease were found from the site www.genecards.org. 74 known proteins were found that are believed to be involved in the pathogenesis of Alzheimer's disease(Table 1).The functional protein sequences in FASTA format for these proteins are collected from NCBI (National Center for Biotechnology Information (http\\www.ncbi.nih.nlm.gov).

2.2 CONSTRUCTION OF PHYLOGENY

All 74 sequences were considered for reconstruction of phylogeny. CLUSTALW (Thompson *et al.,* 1994) was employed for the initial multiple sequence alignment, which calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Based on these results, the scores table and phylogenetic tree that show the distance between the protein sequences was constructed (Figure 1).

2.3 PHYSICOCHEMICAL CHARACTERIZATION

Calculation of physiochemical properties of proteins by traditional experimental methods besides being expensive, is time consuming and cumbersome. We have developed a Java based tool which is used for predicting various physical and chemical properties which may be useful in enhancing our knowledge for experiment design. Physiochemical properties like Length, Number of charged amino acids, Number of hydrophobic amino acids, Number of aliphatic amino acids, Number of aromatic amino acids, Number of polar amino acids for these sequences (Table 1). Amino acid composition of the protein sequences can reveal their nature; hence, amino acid composition was also computed.

S.No	Protein Name	Ac. No.	Length	% of Charged	% of Hydro Phobicity	% of Aliphatic	% of Aromatic	% of Polar
1	A2M	AAH26246	353	25	77	24	10	55
2	ABCB1	AA130425	1280	23	76	24	11	48
3	ACHE	AAH36813	546	20	77	23	12	43
4	AD7C-NTP	AAC08737	375	18	80	21	15	48
5	APLP1	AAH12889	650	28	68	19	08	51
6	APOB	AAH51278	825	26	75	24	08	52
7	APOD	AAH07402	189	20	75	24	11	50
8	APOE	BBA96080	63	28	59	21	11	67
9	APP	AAH65529	751	32	68	19	10	58
10	BACE1	AAH36084	501	21	76	23	12	49
11	ABCA2	AAH08755	867	21	77	23	12	49
12	ABAD	AAH08708	252	17	78	24	06	40
13	BCHE	AAH08396	64	28	70	20	09	65
14	BDNF	AAH29795	247	28	74	19	09	57
15	CASP6	AAH00305	293	30	74	19	14	56
16	CCK	AAHO8283	115	25	70	17	07	50
17	CCR5	AABO9551	215	14	87	29	17	42
18	CD36	CCA83662	472	20	76	25	12	50
19	CD40LG	CAI42902	240	22	75	22	12	54
20	CDH1	AAY68225	882	24	72	22	09	53
21	CDK5	CAG33322	292	29	74	23	11	52
22	CETP	AAB59388	425	20	78	27	11	50
23	CFTR	NP_000483	1480	23	76	26	11	51
24	CHAT	AAI30618	630	24	75	22	10	53
25	CHRNA7	AAH37571	321	21	80	26	11	45
26	CLU	AAH19588	449	30	68	19	11	61
27	CSF1	AAH21117	554	23	73	18	07	54
28	CSNK1D	AAH15775	409	30	72	18	11	53
29	CTNNA3	AAH65819	516	28	71	25	06	54
30	CTSD	CAG33228	412	19	81	24	10	48
31	CYCS	AAH67222	105	32	80	16	11	55
32	CYP19	AAH56258	359	26	78	26	11	50
33	DBN1	AAH07567	649	31	66	14	07	55
34	DCN	AAH05322	359	23	75	25	09	52
35	DSCR1	AAH02864	197	29	74	17	15	57
36	ESR1	CAI42285	595	24	76	20	10	52
37	ESR2	AAH24181	323	25	77	17	11	58
38	FGF2	NP_001997	288	28	74	13	07	48
39	FN1	AAH16875	268	26	69	12	13	65
40	GAL	AAH30241	123	27	76	21	08	48
41	GLUL	AAH51726	373	27	73	15	13	55
42	GSK3B	AAM88578	74	20	70	14	08	57
43	HTR6	AAH74996	440	13	83	25	09	39
44	IGF1R	AAH10607	55	13	87	25	11	44
45	FE65	AAH10854	708	25	70	17	07	50
46	IL18	CAG46798	193	31	67	20	10	58
47	LRP1	AAH21204	439	23	74	21	09	55
48	MAPK1	AAH99905	360	28	73	24	13	53
49	MAPT	AAH0558	352	29	78	14	05	55
50	NCSTN	AAH47621	689	21	75	22	11	53
51	PIN1	AAH31971	45	24	73	13	13	53

52	PLAU	AAH13575	431	25	76	18	12	59
53	PNMT	NP_002677	282	23	74	21	10	48
54	MCP1	AAO75526	25	04	92	36	04	24
55	NP1	AAK61283	473	23	73	21	09	49
56	NgR	ABC69293	600	24	75	22	10	53
57	PAT1	AAC83973	585	28	74	23	15	56
58	IVIg	CAC29069	110	16	80	17	11	54
59	LPL	CAG3335	475	26	77	21	13	55
60	PSEN1	AAH11729	463	21	77	27	13	48
61	PSEN2	CAH73110	448	19	80	27	12	45
62	S100B	AAH01766	92	40	67	20	14	58
63	SNCA	AAI08276	140	28	76	18	05	48
65	STH	AI30322	128	16	76	19	06	52
66	UBB	AAH3899	229	30	70	26	05	55
67	VEGF	AAA35789	191	31	70	16	13	63
68	PRND	AAH43644	176	24	75	21	14	54
69	PARP1	AAH14206	250	42	78	13	09	63
70	MAPK10	AAH35057	461	26	76	22	11	55
71	MAPK14	AAH31574	360	28	73	24	12	53
72	IL1RAPL2	AA10478	686	28	75	24	11	56
73	IL2RA	CAI41071	200	23	73	16	07	60
74	IDE	CAI132670	1019	29	73	22	14	54

Table 1. Table showing genes/proteins that have been studied in the present study, which are believed to be involved in Alzheimer's disease and Percentages of Charged, Hydrophobicity, Aliphatic, Aromatic and polar behavior in amino acids of AD proteins.

Phylogenetic tree

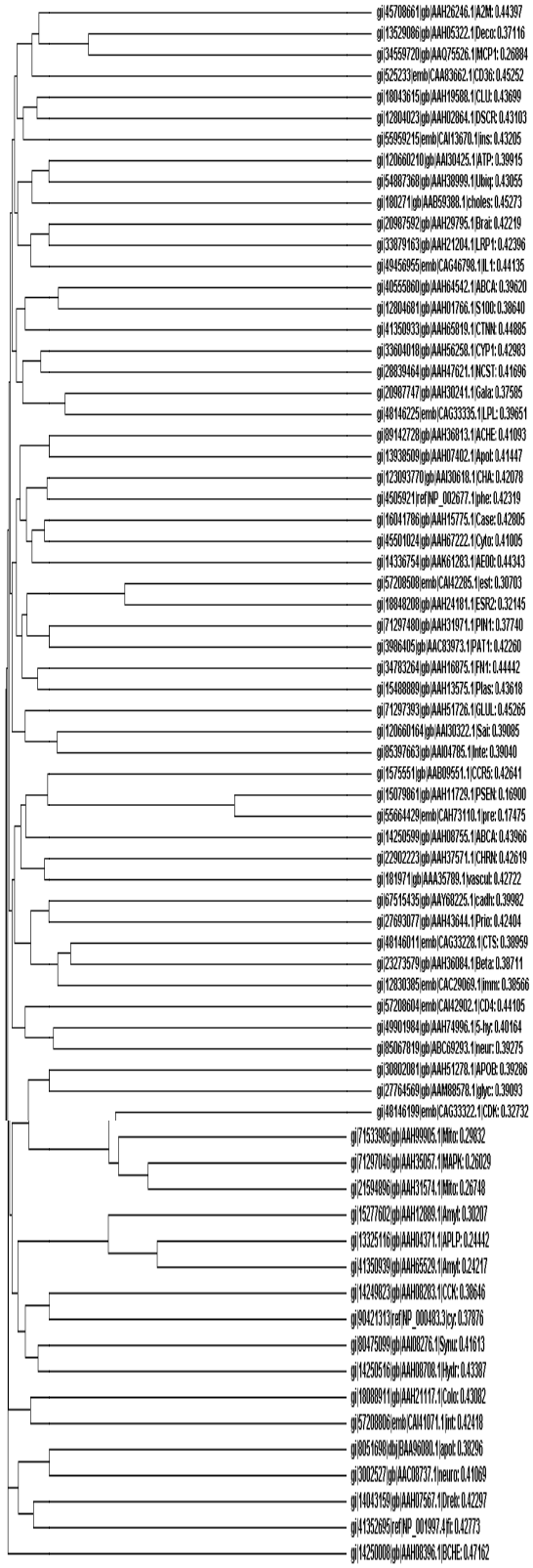
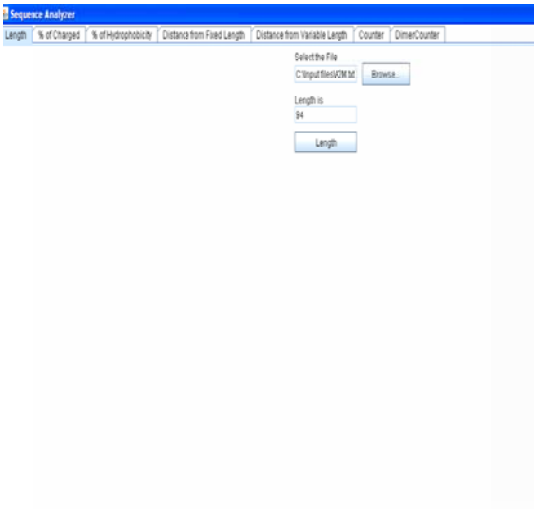
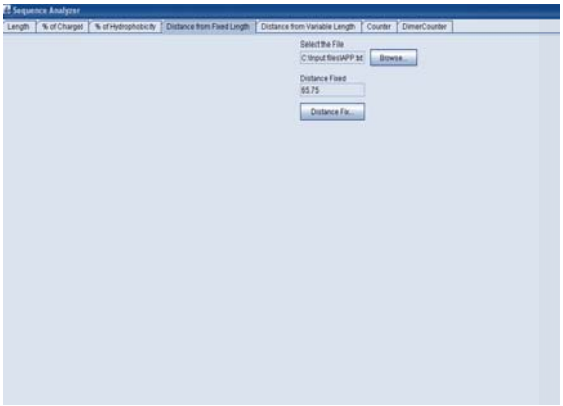
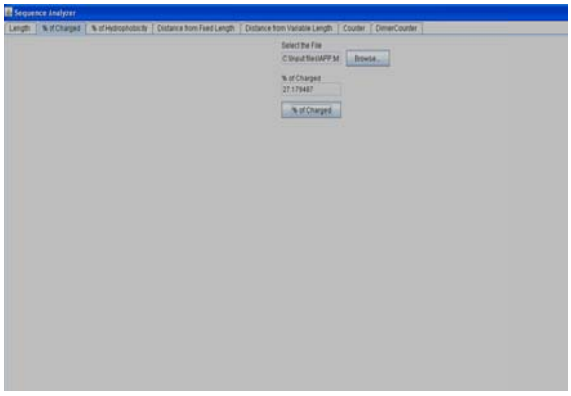


Figure 1. The phylogenetic tree that was constructed based on the alignment score of all the protein sequences involved in Alzheimer's disease. A high degree of homology was noted between presenilin1, presenilin 2, Amyloid beta (A4) precursor protein.







3. RESULTS & DISCUSSION:

The bioinformatics analysis revealed three important proteins out of 74 proteins that are key Pathological proteins in the evolution of Alzheimer's disease. The present bioinformatics study revealed that the proteins: presenilin-1 (PS-1), presenilin-2 (PS-2), and amyloid precursor protein (APP) play a significant role in the pathogenesis of Alzheimer's disease (Figure1). The present study identified the common and different features between the proteins have a greater number of proteins with more charged amino acids whereas some proteins have been observed to have a greater number of hydrophobic proteins. Despite the differences in intrinsic compositional characteristics between the proteins from the different proteins .The study also identified certain common characteristics. Amyloid beta is the major component of amyloid plaques characterizing Alzheimer's disease. The beta amyloid (Aβ) protein is a key molecule in the pathogenesis of Alzheimer's disease (AD). An increased production of Aβ results in neurodegeneration and ultimately dementia through a cascade of events (Giuseppe Verdile et al, 2004**)**.Amyloid beta accumulation can be affected by numerous factors including increased rates of its production and/or impaired clearance. Insulin degrading enzyme is responsible for the degradation and clearance of amyloid beta in the brain (Edland SD, 2004). Amyloid peptides are chains of 40 to 42 amino acid residues. Due to their unique chemical architecture, consisting of water-loving and water-avoiding amino acid sequences, Alzheimer's peptides "self-assemble" to form tangles of fibrils in the brains of persons with Alzheimer's disease (P. Thiyagarajan et al, 2003).

Presenilins act as catalytic subunit of gamma secretase. Presenilins, the causative molecules of FAD, are transmembrane proteins localized predominantly in the ER and Golgi apparatus. Presenillins are thought to be involved in intramembrane proteolysis mediated by their gamma secretase activities. In addition, presenilins interact with FKBP38 (human FK506-binding protein 38) and form macromolecular complexes together with anti-apoptotic Bcl-2, thus it may regulate the apoptotic cell death (Wang HQ et al, 2005).Presenilins and their interacting proteins play a major role in the generation of A-beta from theamyloid precursor protein (APP). Three proteins nicastrin, aph-1 and pen-2 interact with presenillins to form a large enzymatic complex known as gamma secretase that cleaves APP to generate Aβ (Verdile G et al, 2007). Amino acid repeats play an important structural role in proteins and are often associated with diseases. However, repeats are not limited to single amino acids, but can include domains repetitions (Andrade MA et al, 2001). beta A4, which comprises up to 43 amino acid residues. It is highly insoluble under physiological conditions and aggregates into filaments that form very dense clusters in vivo and in vitro. Based on a beta A4 prototype sequence spanning residues 10 to 42 or 43, we have designed analogues in which hydrophobic amino acid residues in position 17 to 20 were substituted by more hydrophilic residues. Depending on the kind of newly introduced amino acids and their position within the sequence, the substitution of only two residues led to variants exhibiting a broad spectrum of different properties. Common to them was a reduced beta-sheet content after solubilization in water and in the solid state. Some of the variants showed significantly reduced amyloidogenicity: although still forming filaments, they did not aggregate into the highly condensed depositions that are typical for amyloid. In addition, they could be solubilized in 200 mM-NaCl and KCl. When mixed with beta A4 peptides bearing the natural sequence, two of the analogues could inhibit the formation of filaments in vitro. These results demonstrate that a well-preserved hydrophobic core around residues 17 to 20 of beta A4 is crucial for the formation of beta-sheet structure and the amyloid properties of beta A4. The introduction of structural alterations within this region may guide the development of reagents for the therapy of Alzheimer's disease (Hilbich C et al, 1992).

4. CONCLUSION:

There is an urgent need for biomarkers to diagnose neurodegenerative disorders early in like AD, when therapy is likely to be most effective, and to monitor responses of patients to new therapies. As research related to this need is currently most advanced for Alzheimer's disease, Shaw LM et al (2007) reviewed the focuses on progress in the development and validation of biomarkers to improve the diagnosis and treatment of AD and related disorders. It is

evident from the preceding discussion that presenilin-1 (PS-1), presenilin-2 (PS-2), and amyloid precursor protein (APP) play a significant role in the pathogenesis of Alzheimer's disease. Missense mutations in APP, PS-1,and PS-2 genes could alter the proteolysis of APP and increase the generation of Aβ42, whose accumulationas diffuse plaques triggers the inflammatory responses due to microglial activation and release of pro inflammatory cytokines. The involvement of inflammatory process in the pathogenesis of Alzheimer's disease is further supported by the observation that inhibition or neutralizing the actions of TNF-α could be of benefit to these patients (Tobinick E et al, 2006; Rosenberg PB, 2006). These evidences and the results of the bioinformatics study reported here strongly suggest that PS-1, PS-2 and APP play a dominant role in the pathogenesis of AD by inducing a pro-inflammatory state.

REFERENCES:

[1]. Allam Appa Rao, Bhremeramba, GR Sridhar, Mathematical analysis of diabetes related proteins having high sequence complexity, Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), 2006.

[2]. Andrade MA, Perez-Iratxeta C, Ponting CP: Protein repeats: structures, functions, and evolution. J Struct Biol 2001, 134(2-3)**:**117-131.

[3]. Edland SD. Insulin-degrading enzyme, apolipoprotein E, and Alzheimer's disease. J Mol Neurosci 2004; 23:213-217.

[4]. Gouchol Pok; Keun Ho Ryu, Co-Occurring Patterns of Amino Acid Physicochemical Properties in Proteins, Web-Age Information Management, (WAIM '08), 2008.

[5]. Giuseppe Verdile, Stephanie Fuller, Craig S. Atwood, Simon M. Laws, Samuel E. Gandy and Ralph N. Martins, The role of beta amyloid in Alzheimer's disease: still a cause of everything or the only one who got caught?, Research, Volume, October 2004, Pages 397-409

[6]. Hilbich C, Kisters-Woike B, Reed J, Masters CL, Beyreuther K., Substitutions of hydrophobic amino acids reduce the amyloidogenicity of Alzheimer's disease beta A4 peptides, J Mol Biol. 1992 Nov 20;228(2):460-73.

[7]. Hutton M, McGowan E. Clearing Tau pathology with amyloid beta immunotherapy—reversible and irreversible stages revealed. Neuron 2004; 43: 293-294

P. Thiyagarajan, New clues found to Alzheimer's disease, logos, vol. 21, no. 2, 2003.

[8]. Rosenberg PB. Cytokine inhibition for treatment of Alzheimer's disease. Med Gen Med. 2006; 8: 24.

[9]. Shaw, L.M., Korecka, M., Clark, C.M., Lee, V.M.-Y.Trojanowski, J.Q., Biomarkers of neurodegeneration for diagnosis and monitoring therapeutics., Nature Reviews Drug Discovery 2007: 6 (4), pp. 295-303

[10]. Tobinick E, Gross H, Weinberger A, Cohen H. TNF alpha modulation for treatment of Alzheimer's disease: a 6-month pilot study. MedGenMed. 2006; 8: 25.

[11]. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-80.

[12]. Verdile G, Gandy SE, Martins RN. The role of presenilin and its interacting proteins in the biogenesis of Alzheimer's beta amyloid. Neurochem Res 2007; 32:609-623.

[13]. Wang HQ, Nakaya Y, Du Z, Yamane T, Shirane M, Kudo T, Takeda M, Takebayashi K, Noda Y, Nakayama KI, Nishimura M. Interaction of presenilins with FKBP38 promotes apoptosis by reducing mitochondrial Bcl-2. Hum Mol Genet 2005; 14: 1889-1902.

AUTHOR PROFILES:

Dr. Kiran Kumar Reddi has received PhD in Computer Science and Engineering from Acharya NagarjunaUniversity, Guntur, Andhra Pradesh, India.He is working as Assistant Professor in the department of Computer Science, Krishna University, Machilipatnam, Andhra Pradesh, India. His research interest includes Bioinformatics, Software Engineering and Network Security. He is a member of professional societies like CSI and ISTE.

Dr. T K Rama Krishna Rao has received his PhD in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. He is working as Professor in the department of Information Technology, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India. His research interest includes Bioinformatics, Software Engineering and Network Security. He is a member of professional societies like CSI and ISTE.

Mr. G.Nageswara Rao has received his M.Tech in Computer Science and Engineering from Andhra University, Waltair, Andhra Pradesh, India. He is working as Professor in the department of Information Technology, Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India. His research interest includes Bioinformatics, Software Engineering and Network Security. He is a member of professional societies like CSI and ISTE.