Aggregation prediction;
A bioinformatics method for studying the effects of missense variations on pathogenicity

Master's Thesis
Bioinformatics Masters Degree Programme,
Institute of Biomedical Technology,
University of Tampere,
Finland.
Percy, N. Hevor
December 2011

ACKNOWLEDGEMENTS

I want to express much gratitude to my supervisor and Group Leader, Prof Mauno Vihinen for his devoted time in helping me with all the difficulties encountered in undertaking this project. My acknowledgements also go to Martti Tolvanen for his guidance and words of encouragement. Special thanks to Ayodeji E. Olatubosun for his corrections and help with my results.

I also want to acknowledge my parents, Mr. and Mrs. Hevor for their financial and moral support during the time of this program. Not forgetting my dear wife Linda Tamatey for her love and support. Thanks to all my siblings and friends who made sure I was in good spirit to get my work done. Finally, I thank God for giving me the wisdom and good health to successfully complete this project.

December, 2011.

Percy, N. Hevor
Bioinformatics Programme,
Institute of Biomedical Technology,
University of Tampere,
Finland.

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE Bioinformatics Masters Degree Programme. Institute of Biomedical Technology, Tampere, Finland.
Author:	Percy, N. Hevor
Title:	Aggregation prediction; A bioinformatics method for studying the effects of missense variations on pathogenicity
Pages:	40 pp + Figures 13 + Tables 3+Appendices 7pp
Supervisor:	Prof. Mauno Vihinen
Reviewers:	Prof. Mauno Vihinen, Csaba Ortutay
Time:	December 2011.

Abstract

Background and Aims: Aggregation has been shown to be an intrinsic property of many proteins including proteins not involved in amyloid diseases. The most common types of protein aggregates are amyloid fibrils and amorphous aggregates characterized by an increase in the level of β-structure. Missense variations have the potential to change the propensity of a property to aggregate. Variation research in recent times has focused on obtaining information about the effects of sequence variations on proteins. Experimental study of the possible disease association of variants is laborious and time-consuming. Computational methods on the other hand give rapid automated results for large amounts of data sets but are less reliable. To use aggregation as mechanism to study the effects of missense variations on pathogenicity, it is important to predict the change in aggregation of proteins upon aggregation. There are several aggregation prediction methods available on the Internet making it difficult to find the best methods. This study evaluates the performance of five widely used aggregation prediction methods. Results from the aggregation prediction can then be used for pathogenicity prediction to determine how they correlate.
Methods: Aggrescan, AmylPred consensus, Average Packing Density, TANGO and Hexapeptide Conformational Energy were the evaluated methods. The methods were tested with a dataset of 365 missense variations. Matthews correlation Coefficient, Sensitivity, Specificity, Accuracy, Precision and Negative Predictive Value were the measures used to evaluate the performance of the prediction methods.
Results: Aggrescan performed best in MCC, accuracy, sensitivity and NPV show that is the best method. Tango performed best in precision (0.92) and specificity (0.95).
Conclusion: From the results, all the methods showed good MCC values of above 0.59. It is easy to conclude that Aggrescan was the best amongst all the five methods followed by Tango. It is on the other hand difficult to recommend a specific method since all the methods depend on physicochemical properties and side chains in β-sheet aggregates making the algorithms in the methods give different results. It is therefore advisable for the end user to know much about the algorithms used before choosing a particular method for prediction.

CONTENTS

ABBREVIATIONS

AD	Alzheimer's disease
AD&FTDMDB	Alzheimer Disease & Frontotemporal Dementia Mutation Database
Aggrescan	For the prediction of "hot spots" of aggregation in polypeptides
ALS	Amyotrophic lateral sclerosis
AmylPred	A web tool for a consensus prediction of amyloidogenic determinants
BSC	Bovine Spongiform Encephalopathies
CJD	Creutzfeldt-Jakob disease
FPR	False positive rate
FTD	Frontotempral dementia
G	Average Packing Density
HD	Huntington's disease
HS	Hot spots
Indels	Insertions and Deletions
MCC	Mathews correlation coefficient
NLM	National Library of Medicine
NPV	Negative predictive value
PD	Parkinson's disease
PON-P	Pathogenic-or-Not Pipeline
SBS	Single Base Substitutions
TANGO	A web based tool for the prediction of protein aggregation
TPR	True positive rate
TSE	Transmissible spongiform encephalopathies
Z	Hexapeptide Conformational Energy
3D	Three-dimensionl

1. Introduction

In recent years, it has become increasingly important to know the effects of missense variation on proteins. This is due to the fact that missense variations have an effect on many biochemical properties including disorder, stability or aggregation propensity of a protein. In most cases, missense variations are known to result in deleterious effects on proteins but there are also few instances where the mutations have positive effects on the protein. In such rare cases, missense variations result in new versions of proteins that help an organism and its future generations adapt better to environmental changes. An example is when a beneficial mutation produces a protein that increases the resistance of the organism to a new strain of bacteria (Genetics Home Reference, 2011).

Missense variations can impair the activity of many enzymes by an increase in protein instability. It may also increase, decrease or have no effect on the propensity of a protein to aggregate. Aggregation has been linked to the pathogenesis of most neurodegenerative diseases such as Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD), amyotrophic lateral sclerosis (ALS) and prion diseases (Fink, 1998).

The toxicity associated with protein aggregates in many neurodegenerative disorders has been attributed to abnormal interactions between misfolded proteins with normal cellular constituents. Protein aggregation narrows the spectrum of relevant polypeptides obtained by recombinant techniques. It reduces the shelf life and increases the immunogenicity of polypeptidic drugs (Ventura and Villaverde, 2006).

Aggregation-prone zones of proteins can be determined experimentally in the laboratory but it is laborious and time consuming due to the large amount of variation data available. Major advances in the prediction of aggregation-prone zones based on sequence analysis using different web based methods have therefore been on the increase. Previous research mainly concentrated on using just one or a few methods in one study (Burke *et al*., 2007; Lappalainen *et al*., 2008; Tavtigian *et al*., 2008; Thusberg and Vihinen, 2006, 2007; Worth *et al*., 2007). In this study, I used the Pathogenic-or-Not Pipeline (PON-P) that provides simultaneous access to five extensive aggregation prediction methods. Combination of these methods has the advantage of the compensation of the weakness of one program by the others resulting in a more reliable prediction.

The number of available aggregation prediction methods is on the increase and each method has it's own algorithms. It is therefore important to know which of the methods are closer to predicting the right aggregation-prone zones. This is therefore the main focus of this study.

1.1 Aim and Objectives

The main aim of this study is to use bioinformatics methods to predict the aggregation propensities of as many mutated proteins as possible and to deduce the usefulness and reliability of the existing methods.

Objectives include:

1. Using web-based methods available at the Pathogenic-Or-Not (PON-P) website, to obtain the aggregation propensities of wild type sequences of proteins and their mutated sequences that are experimentally proven to have an increase, decrease or no effect in aggregation after mutation.
2. Measurement of the performance of the aggregation prediction methods using six measures.
3. Finding out how well the methods work and how it can help the end user select the best method.

1.2 Significance of the study

This work is part of the Pathogenic or Not Project (Thusberg and Vihinen, 2009). Focusing on Aggregation as one of the sequence-based analysis for the prediction of the effect of missense variations on the pathogenicity of proteins. There are many bioinformatics methods, most of which are freely available online that can be used to predict the aggregation propensity of proteins. The methods used for this study are found in the PON-P portal. They were used to analyze 365 missense variations from 58 different proteins.

Missense variations have the ability to affect protein posttranslational modifications leading to diseases. However due to the redundancies of cellular pathways, it has been found out that not all the mutations are pathogenic. The results from the aggregation prediction can be used to determine how it correlates with the pathogenicity of the proteins after the introduction of a variation.

This study focuses on the comparison of the available bioinformatics aggregation prediction methods, which will help the end user select the best methods.

2. Review of Literature

A recent advance in experimental work has helped to identify the residues within a protein sequence that promote ordered aggregation and amyloid formation (Fernandez-Escamilla *et al.,* 2004). Research indicates that aggregation is generally favored by mutations. Protein aggregation and the formation of highly insoluble amyloid structures is linked with a wide range of debilitating human conditions (Fink, 1998). Sequence-dependent methods of prediction of aggregation-prone zones in a protein were developed based on the observation of the experimental work using statistical mechanics algorithms. (Fernandez-Escamilla *et al.,* 2004)

The application of high-throughput sequencing methods has increased the available pool of data for identified variants in the human genome increasing the difficulty to identify disease-causing mutations. Research to computationally determine whether the mutation is pathogenic or benign has therefore become a topic of growing interest in recent years. This could significantly help to target disease-causing mutations by helping in the selection and prioritization of likely candidates from the plethora of data on gene defects (Thusberg *et al.,* 2011).

2.1 Variations

Variations introduce permanent changes in the genomic sequence of a protein. The changes may come as a result of endogenous processes or interaction with exogenous agents. Endogenous processes include DNA replication errors, the intrinsic instability of certain DNA bases or from attack by free radicals generated during metabolism. Exogenous agents such as ionizing radiation, UV radiation and chemical carcinogens, upon interaction with proteins can also result in changes in the genomic sequence (Bertram, 2000). Variations play a major role in the determination of changes in the characteristics of populations across multiple generations.

There are different types of variations in proteins. These include insertions and deletions (Indels), Single Base Substitutions (SBS), duplications and translocations. Indels result in frameshift variation that disrupts the reading frame provided the number of nucleotides inserted or deleted is not a multiple of three. On the other hand, Insertion or deletion of three nucleotides, results in an extra or a missing amino acid in the final protein (Http://www.nature.com/scitable/definition/frameshift-mutation-frame-shift-mutation-frameshift-203).

SBS also termed as Point variations involve the replacement of a single base by another. If the substitution involves the replacement of one purine [A or G] or pyrimidine [C or T] by the other, the substitution is termed a Transition. If on the other hand a purine is replaced by a pyrimidine or vice-versa, the substitution is called a Transversion.

SBS are categorized into nonsense variations, silent variations and missense variations. A nonsense variation is a point variation resulting in a premature stop codon due to a base change in the DNA that prematurely stops the translation of mRNA and truncates the protein rendering it nonfunctional. Silent variations are point variations that occur in non-coding regions or within an exon without introducing a functional change in the protein because it does not result in a change to the amino acid sequence of a protein.

This study concentrates on missense variations. It seeks to evaluate the effects of missense variations on a protein using different bioinformatics methods of aggregation.

2.1.1 Missense variation

Missense variation (*Figure. 1*) is a nonsynonymous mutation, which involves the substitution of a single base in the coding region of a DNA for another resulting in the substitution of one amino acid in a polypeptide for another.

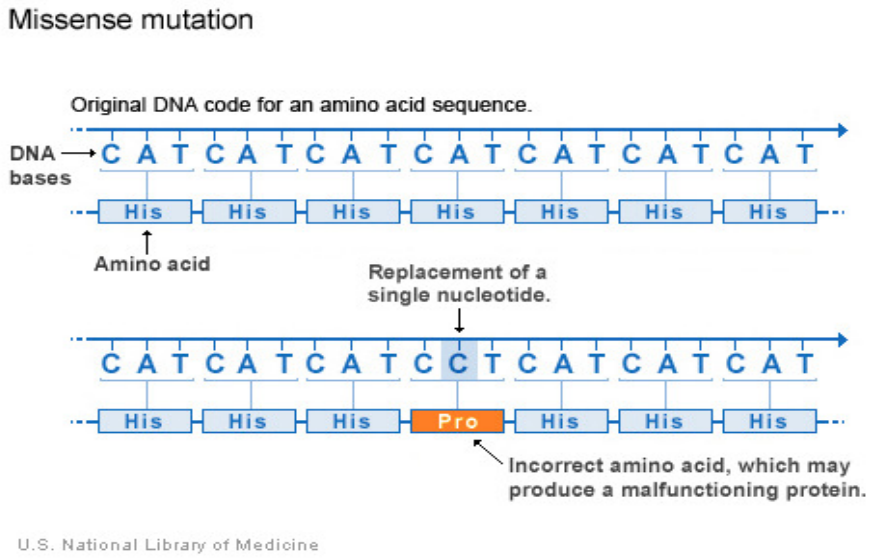

***Figure1**:* Missense variation. [Http://ghr.nlm.nih.gov/handbook/illustrations/missense]

Some missense variations alter a gene's DNA base sequence but do not change the function of the protein made by the gene. This is as a result of multiple redundancies of cellular pathways (Thusberg and Vihinen, 2009).

Each cell has a number of pathways through which enzymes recognize and repair mistakes in DNA.

A very small percentage of all mutations actually have a positive effect. Positive mutations lead to the formation of novel proteins with diverse biological functions that help an organism and its future generations better adapt to changes in their environment (Hamadrakas *et al*., 2007).

Most missense variations on the other hand are deleterious. They may lead to significant changes in the protein structural properties, causing abnormal folding, structural instability, or aggregation of the protein (Thusberg and Vihinen, 2009). Studies have shown that there are over 20 known familial diseases caused by single point mutations that result in an increase in the probability of aggregation and neurodegeneration (Gerum *et al.,* 2010).

Missense variations like prion proteins have been associated with a range of familial diseases. The prion protein (PrPc), a normal protein found in the membranes, misfolds into a pathogenic form PrPSc, a highly ordered fibrillar aggregate. Prion aggregation can take place both extracellularly and intracellularly (Ross and Poireir, 2004). PrPSc renders the protein resistant to proteases, resulting in the development of different neurodegenerative diseases. These diseases termed transmissible spongiform encephalopathies (TSEs) produce lethal decline of cognitive and motor function (Horwich and Weissman, 2007). There are many varied types of TSEs. The most common examples are Creutzfeldt-Jakob disease (CJD) in humans and Bovine Spongiform Encephalopathies (BSC also known as mad cow disease) in cattle.

Missense variations change the properties of the protein increasing the tendency of the protein to aggregate. It has been suggested that the composition and the primary structure of a protein plays a major role in the determination of the aggregation propensity of the protein and even small changes may have a considerable effect in the solubility of the protein (Thusberg and Vihinen, 2009).

2.2 Aggregation and amyloid fibril formation

Protein aggregation is the conversion of peptides and proteins into insoluble fibrillar aggregates. It is the ability of highly soluble proteins in biological fluids to gradually misfold into insoluble filamentous polymers with β-pleated sheet conformation (Selkoe, 2003; Ross and Poireir, 2004). Protein aggregates can be divided into various types, including disordered or 'amorphous' aggregates and amyloid formation.

Amyloid fibrils are the most characteristic types of aggregation. They have a more compact structure than other types of aggregation, which makes it difficult to access proteases and breakdown the amyloid (Maurer-Stroh *et al.,* 2010). Amyloid fibrils are formed when well-organized fibrillar aggregates are deposited extracellularly. It involves the polymerization of abnormal states of normally soluble proteins or peptides (Kelly, 1998). Formation of amyloid-like fibrils plays a key role in about 40 human protein deposition diseases. These include Alzheimer's disease, type II diabetes, prion diseases and Parkinson's disease, collectively called amyloidoses.

These diseases result in neurodegenerative, metabolic and systematic symptoms with the deposition of proteinaceous aggregates in various tissue types (Aisenbrey. *et al.,* 2008).

In the amyloid diseases, a diverse group of normally soluble proteins self-assemble to form insoluble fibrils also known as precursor proteins (Jimenez *et al.,* 1999). In their native soluble forms, these precursor proteins do not have a general sequence or a three dimensional (3D) structural homology. They all assemble into a cross-β-fiber structure; with β -strands perpendicular and β -sheets parallel to the fibril axis (Sunde *et al.,* 1997) and can bind Congo red with characteristic birefringence (Klunk *et al.,* 1999).

It was once thought that relatively few proteins have the propensity to aggregate but recent data suggest that many soluble proteins can, under destabilizing circumstances, undergo this conversion *in vitro* (Selkoe, 2003). Amyloid fibril formation is therefore considered a common property of proteins with varying individual propensities as a result of sequence and environmental conditions (Monsellier and Chiti, 2007).

Research supports the concept that the occurrence of amyloidogenic and intrinsically disordered regions has similar factors in different peptides and proteins. One major problem is the recognition of these factors that influence protein conformational changes and misfolding, the solution to which will be important in finding effective treatments for amyloid illnesses (Galzitskaya*, et al,* 2006*)*

Proteins have evolved many sequence and structural adaptations to counteract their natural tendency to aggregate into amyloid-like fibrils (Monsellier and Chiti, 2007). Molecular evolution has acted on protein sequences to finely modulate their aggregation propensities depending on different parameters related to their *in vivo* environment. This together with cellular control mechanisms protects proteins from aggregation during their lifetime (Monsellier and Chiti, 2007) Research has shown that, glycine residues appear to be evolutionarily conserved in their ability to inhibit aggregation (Parrini *et al.,* 2005). There has been a suggestion that, structurally related proteins have a positive evolutionary pressure to maintain glycine residues at specific positions in order to preserve their overall architecture (Branden and Tooze, 1999).

Fibril formation makes normal proteins toxic (Bucciantini *et al.,* 2004). This may be the case of peripheral amyloidoses, which results in physical disruption of normal tissues function due to massive accumulation of amyloid fibers (Dobson, 2003). Previous work suggests that mature fibrils are substantially harmless compared to highly toxic pre-fibrillar aggregates (Walsh *et al.,* 2002; Stefani and Dobson, 2003; Kayed *et al.,* 2003). Leading to the proposition that the pre-fibrillar assemblies share basic structural features that, at least in most cases, seem to underlie common biochemical mechanisms of cytotoxicity (Stefani and Dobson, 2003; Kayed *et al.,* 2003; Bucciantini *et al.,* 2002; Bucciantini *et al.,* 2004, Bucciantini *et al.,* 2005).

However, an increase in evidence suggests that amyloid formation may result in a protective mechanism, which especially in the case of the neurodegenerative amyloidoses, acts as to sequester misfolded polypeptides that would otherwise dwell in more toxic, and more highly interactive, oligomeric species (Bryan *et al.,* 2011). Chorion, the major component of silkmoth eggshell is an example of a natural amyloid that protects the silkmoth oocyte and embryo (Iconomidou *et al.,* 2000).

The propensity to form amyloid fibrils can vary between different sequences even though the ability to form amyloid fibrils seems to be generic (Dobson, 2003). There is a high correlation between the relative aggregation rates for a wide range of peptides and proteins and physicochemical features of the molecules such as charge, secondary-structure propensities and hydrophobicity (Chiti *et al.,* 2003). In the determination of aggregation propensity of proteins, it is not all the regions of a polypeptide are of equal importance. There are specific segments that can nucleate when exposed to solvent, giving an indication of sequence dependence of aggregation properties (Conchillo-Solé *et al.,* 2007).

Very short stretches of specific amino acids modulate aggregation by acting as facilitators or inhibitors of amyloid fibril formation (Ivanova *et. al.,* 2004; Ventura *et al.,* 2004). The presence of these relevant regions, also termed as aggregation "hot spots" (HS) (Conchillo-Solé *et al.,* 2007) has been described in most of the peptides and proteins underlying neurodegenerative and systemic amyloidogenic disorders (Chiti and Dobson, 2006). Some short peptides possess the same amyloid properties as full-length proteins (Balbirinie *et al.,* 2001; Tendis *et al.,* 2000). Averagely however, long proteins, have less intense aggregation peaks than short ones (Monsellier *et al.*, 2007). Proteins with different subcellular localizations were found to have different aggregation propensities due to their different structures and cellular microenvironments (Monsellier and Chiti, 2007).

Relevant polypeptides obtained by recombinant techniques are one major area in protein production that has been affected by protein aggregation. It narrows the spectrum of relevant polypeptides obtained (Ventura and Villaverde, 2006). In pharmaceutical research, aggregation can increase the cost or time needed for the production of antibodies and small molecules that are developed when over-expressed .It also reduces the shelf life and increase the immunogenicity of polypeptidic drugs.

The identification of amyloid aggregates has therefore become a globally critical research topic, which involves collective effort to address this problem by developing new therapies that interfere with the early stages of a proteins ability to form aggregates. This research topic could further be used for developing drugs that counter amyloid formation (Maurer-Stroh *et al.,* 2010).

Development of algorithms capable of predicting aggregation parameters of unstructured polypeptides directly from their amino acid sequence has also gained much progress based on the idea of the generality of amyloid fibril formation.

These algorithms have the ability to predict many aggregation related parameters, including the aggregation propensity of a peptide chains, aggregation-prone regions and the effect of mutations on the aggregation behavior (Monsellier *et al.,* 2008).

2.3 Bioinformatics methods for the analysis of variations

Bioinformatics methods have become very important tools for the analysis of the effects of missense variations. This is because conducting experimental analysis is laborious and time-consuming. Although, clearly, prediction tools cannot entirely replace experimental work, they might contribute in locating potential regions of interest for further experimental studies. The methods help in the analysis of a lot of data simultaneously (Kimon *et al.,* 2009).

Emerging trend in mutation analysis is to utilize a more extensive set of prediction methods in order to attain more reliable results which is contrary to past when research mainly concentrated on using just one or a few methods in one study (Burke *et al.,* 2007; Lappalainen *et al.,* 2008; Tavtigian *et al.,* 2008a,b; Thusberg and Vihinen, 2006,2007; Worth *et al.,* 2007).

Mutation databases serve as the basis for the prediction methods, providing the data for the analysis (Thusberg and Vihinen, 2008). The methods for the analyzing the effects of missense variations are divided into sequence-based and structure based. Some of these methods however do overlap. Sequence-based analysis methods include; cellular localization and aggregation. Structure-based analysis methods are electrostatic changes, steric effects and changes in inter-residue contacts.

Some of the methods that are considered as both sequence- and structure-based are disorder, functional effects and stability.

In this study, I focused on aggregation as a sequence-based analysis method. Many high-throughput methods have been developed for the prediction of aggregation propensities of proteins from protein primary structure. Some of these methods include Agrescan, AmylPred, BetaScan and BetaWrapPro. These methods are included in the methods used by the Pathogenic-Or-Not Pipeline (PON-P) for aggregation Prediction.

2.4 PON-P, Pathogenic-or-Not Pipeline

PON-P (freely available at http://bioinf.uta.fi/PON-P/) is a service developed by Vihinen *et al*. that provide simultaneous access to numerous methods for the prediction of the effects and the consequences of variations in proteins. PON-P combines methods from different categories. These include stability change prediction, aggregation prediction, disorder prediction, tolerance prediction etc.

Combinations of these methods give a more reliable prediction by compensating the weakness of one program by the others. The pipeline can be used even for larger sets of variations. Results contain interpretation of the output of the individual predictions and overall summary of the pathogenicity (Thusberg and Vihinen, 2008).

Aggregated or Not-Pipeline (*Figure 2* below) is part of PON-P but focuses on only Aggregation methods. It simultaneously submits the input data provided by the user to five aggregation methods.

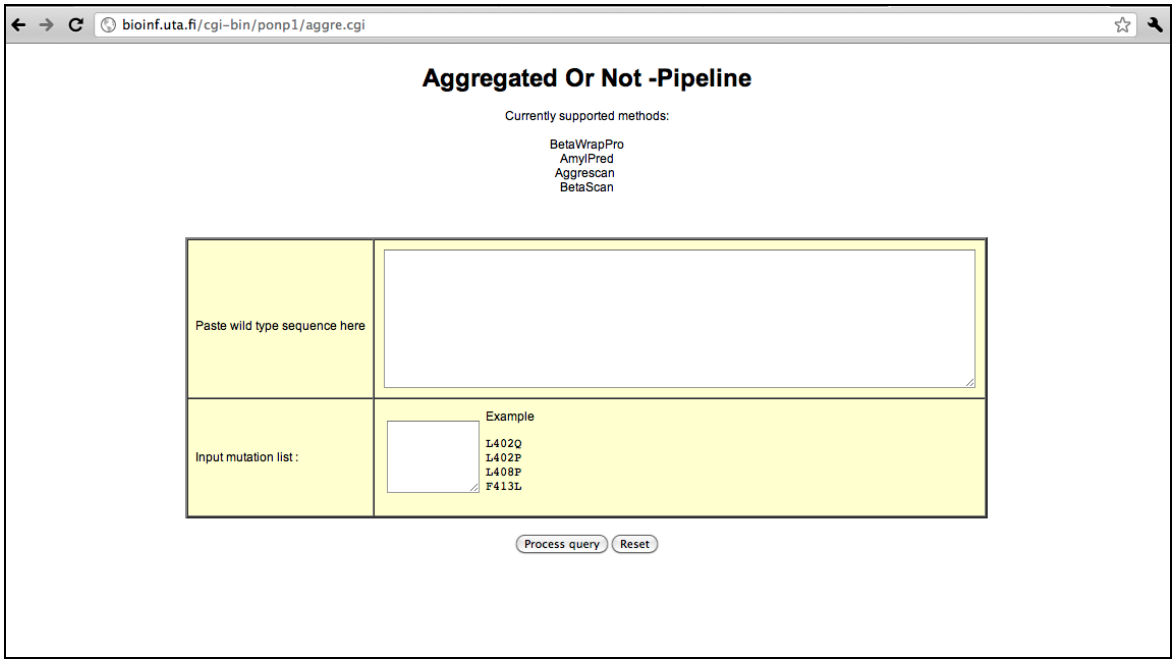

***Figure 2**:* Aggregated or Not –Pipeline. (*www.bioinf.uta.fi/cgi-bin/ponp1/aggre.cgi*)

3. Materials and Methods

3.1 Data set

All the data set used in this work was downloaded online (available in the *appendix*). I built a positive data set (aggregation increasing variations) of 319 missense variations, negative data set (aggregation decreasing variations) of 30 missense variations and a neutral data set (variations that have no effect on aggregation) of 16 missense variations all from a total of 58 different proteins. A single protein may have more than one variation in the protein sequence. This accounts for the total number of 365 variations. Some proteins may have an increase, decrease or no effect on aggregation depending on the type of missense variation that it is subjected to. Variations that occur outside the 'hotspot' area of the protein sequences normally have no effect on the aggregation propensity of the protein.

The proteins used in this study include proteases and hydrolases with biological processes such as apoptosis and cell adhesion. Presenilin1 (PSEN1) and Presenilin2 (PSEN2) are examples of such proteins. Some of the proteins including Transthyretin and Amyloid precursor proteins (APP) are transport proteins. Transthyretin transports thyroxine from the bloodstream to the brains and APP participates in the reverse transport of cholesterol from tissues to the liver for excretion. There are also fibrous and filamentous proteins such as Lamins and Desmins respectively. Lamins are thought to provide a framework for the nuclear envelope and may also interact with chromatin. Desmins are intermediates found in muscle cells.

Defects in most of the proteins used results in amyloidosis and neurodegenerative disorders such as Alzheimer disease, Atherosclerosis, Cardiomyopathy, Neuropathy and Myofibrillar myopathy

The Ovid Medline database and The Alzheimer Disease & Frontotemporal Dementia Mutation Database (AD&FTDMDB) were the major databases for the collection of the data for this project. A search was made for missense variations in proteins that affected their propensities to aggregate, reported in literature and at scientific meetings.

3.1.1 Ovid Medline

MEDLINE is one of the many databases hosted by the Ovid medical information company. MEDLINE, an index used to find articles published in biomedical journals, is produced by the USA National Library of Medicine, NLM (http://www.nlm.nih.gov) and it includes citations with abstracts from approximately 5,400 biomedical dental and nursing journals. NLM provides free access to MEDLINE through PubMed (http://pubmed.gov/). There are 2 search options in MEDLINE.

One from the year 1948 to the present with Daily Update which yields more results and the other from 2007 to the current Update. The default-searching field is 'Advanced search' but there is the option to do a Basic search as well. To search effectively, it is important to break the search terms into concepts and search each concept one at a time before combining the results of the searches. The different concepts are in the search history menu and it is possible to find articles that relate to all the concepts .In the search history menu, all the results of the different concepts are selected and the selections are combined with an 'and' button. The results targeted to all concepts of the search topic are displayed by the click of the 'display' button to reveal a list of all the citations. It is possible to click the 'view abstracts' tab to reveal summary of the articles. There is an option to click 'complete reference' to find more information including how the article has been indexed in Medline. This can help find other related subject headings to search. Clicking the 'get it' button retrieves the article.

3.1.2 Alzheimer Disease & Frontotemporal Dementia Mutation Database (AD&FTDMDB)

The AD&FTDMDB (freely available at http://www.molgen.ua.ac.be/admutations) aims at collecting all known mutations and non-pathogenic coding variations in the genes related to Alzheimer Disease (AD) and frontotempral dementia (FTD). The website was launched as a locus-specific database in September 1999 based on the guidelines of the Human Genome Variation Society (Horaitis et al., 2007). In 2007, a link to the UCSC human genome browser was made in collaboration with PhenCode. There is continuous update of the database and it contains mutations reported in the literature and at scientific meetings, and unpublished mutations directly submitted to the database. To date, the database contains mutations in the genes encoding the Amyloid Beta Precursor Protein (*APP*), Presenilin 1 (*PSEN1*), Presenilin 2 (*PSEN2*), Chromatin Modifying Protein 2B (*CHMP2B*), fusion (involved in t(12;16) in malignant liposarcoma) (*FUS*), Granulin (*GRN*), Microtubule Associated Protein Tau (*MAPT*), TAR DNA binding protein (*TARDBP*) and Valosin-containing Protein (*VCP*) and holds 415 different mutations observed in 1027 patients or families. (Cruts M. *et al*., 1998, Rademakers R. *et al*., 2004, Gijselinck I. *et al*., 2008, Theuns J. *et al*., 2006)

3.2 Aggregation prediction methods

The fact that prediction tools cannot entirely replace experimental work (Kimon *et al.,* 2009) makes it important to know which of the available predictions methods is more reliable in locating potential regions of interest for further experimental studies. It is normally a major interest to be secured that the prediction tool uses algorithms that will be able to perform well on novel data that was not included in the process of constructing the algorithm (Baldi *et al.,* 2000).

Knowledge of the best methods will further help the end user to choose the best method for a particular task.

This study uses some of the available web-based methods in predicting aggregation propensity of mutant proteins and finding their correlation with available experimental data. The web-based methods used include, AGGRESCAN, AmylPred, Hexapeptide Conformational Energy (Z), Tango and Average Packing Density (G) all of which are part of PON-P's aggregated or not pipeline (*www.bioinf.uta.fi/cgi-bin/ponp1/aggre.cgi*).

Table 1: Aggregation prediction methods

Methods	**Web address**	**Predicts**	**Reference**
Aggrescan	http://bioinf.uab.es/aggrescan/	Aggregation prone segments	Oscar et al., 2007
AmylPred	http://biophysics.biol.uoa.gr/AMYPRED	Features related to the formation of amyloid fibrils	Kimon K. F. *et al.,* 2009
Average Packing Density (G)	http://biophysics.biol.uoa.gr/AMYPRED	Amyloidogenic and disordered regions	Galzitskaya OV *et. al.,* 2006
Tango	http://tango.embl.de/ http://biophysics.biol.uoa.gr/AMYPRED	Protein aggregation	Fernandez-Escamilla AM *et. al.,* 2004
Hexapeptide Conformational Energy (Z)	http://biophysics.biol.uoa.gr/AMYPRED	The amyloid fibril-forming segment of proteins	Z. Zhang *et. al.,* 2007

3.2.1 AGGRESCAN

A web-based software (http://bioinf.uab.es/aggrescan/) developed for the prediction of aggregation-prone segments in protein sequences, the analysis of the effect of mutations on protein aggregation propensities and the comparison of the aggregation properties of different proteins or protein sets (Conchillo-Solé *et al.,* 2007). The software is based on an aggregation-propensity scale for the 20 natural amino acids derived from *in vivo* experiments and on the assumption that short and specific sequence stretches are responsible for protein aggregation.

Relative experimental aggregation propensities, for each of the 20 natural amino acids, were initially derived from the intracellular aggregation of mutants, performing single-point mutations at the central position (19) of the central hydrophobic cluster, comprising residues 17-21, of amyloid Aβ1-42 Alzheimer's peptide.

Then, a value is assigned to each residue of a given polypeptide sequence, which is taken from the table giving the relative experimental (*in vivo*) aggregation propensities of the 20 natural amino acids (a3v). Next, calculations are based on the sliding-window averaging technique: A sliding window of a given length is chosen and the program calculates the average of a3v's over the sliding window and assigns it to the central residue of the window. This average is called a4v. A plot of a4v over the entire sequence defines the aggregation profile (AP) of the polypeptide. The "hot spot" threshold (HST) was defined as the average of the a3v of the 20 natural amino acids weighted by their frequencies in the SwissProt database. A segment of the polypeptide sequence is considered as a putative aggregation "hot spot" (HS) if there are 5 or more consecutive residues with an a4v larger than the HST and none of them is a proline (aggregation breaker). Several other parameters are calculated and reported, like: the average a4v in each "hot spot" (a4vAHS), the area of the aggregation profile above the "hot spot" threshold (AAT), the total area (TA, the HST being the zero axis) and the area above the HST of each profile peak identified as a "hot spot". Normalized sequence sum for 100 residues (Na4vSS) is calculated. The change in normalized a4v sum (Na4vSS) and Total Area (TA) are obvious indicators of changes in aggregation properties of the complete sequence due to point mutation. Negative Na4vSS values suggest overall low aggregation propensity and vice versa.

In the AGGRESCAN output, the sequence stretches with highest predicted aggregation propensity are shown in red in the peptide sequence column (*Figure 3*) and appear as peaks in the Profile plots (*Figure 4*). The HS can be ranked according to their peak area (HSA) or normalized peak area (NHSA).

Aggrescan results

	wildtype	Average over all sequences
Sequence Name:		
Graphics:		
a3v Sequence Average (a3vSA):	-0.177	-0.177
Number of Hot Spots (nHS):	8	8.000
Normalized nHS for 100 residues (NnHS):	1.691	1.691
Area of the profile Above Threshold (AAT):	12.203	12.203
Total Hot Spot Area (THSA):	6.253	6.253
Total Area (TA):	-74.969	-74.969
AAT per residue (AATr):	0.026	0.026
THSA per residue (THSAr):	0.013	0.013
Normalized a4v Sequence Sum for 100 residues (Na4vSS):	-17.9	-17.900
		Sorted by Na4vSS
		wildtype -17.90

#	AA	a4v	HSA	NHSA	a4vAHS
1	M	-0.172	0.000	0.000	0.000
2	T	-0.172	0.000	0.000	0.000
3	T	-0.172	0.000	0.000	0.000
4	C	-0.172	0.000	0.000	0.000
5	S	-0.172	0.000	0.000	0.000
6	R	-0.051	0.000	0.000	0.000
7	Q	-0.161	0.000	0.000	0.000
8	F	-0.173	0.000	0.000	0.000
9	T	-0.076	0.000	0.000	0.000
10	S	-0.215	0.000	0.000	0.000
11	S	-0.237	0.000	0.000	0.000
12	S	-0.151	0.000	0.000	0.000
13	S	0.016	0.036	0.000	0.016
14	M	-0.192	0.000	0.000	0.000
15	K	-0.012	0.008	0.000	-0.012
16	G	-0.034	0.000	0.000	0.000
17	S	-0.056	0.000	0.000	0.000
18	C	-0.078	0.000	0.000	0.000
19	G	0.114	0.254	0.051	0.031
20	I	-0.017	0.254	0.051	0.031
21	G	0.019	0.254	0.051	0.031
22	G	0.019	0.254	0.051	0.031
23	G	0.019	0.254	0.051	0.031
24	I	-0.063	0.000	0.000	0.000
25	G	-0.127	0.000	0.000	0.000
26	G	-0.127	0.000	0.000	0.000
27	G	-0.105	0.000	0.000	0.000

***Figure 3**:* Aggrescan results of Keratin type I cytoskeletal mutant (R127P).

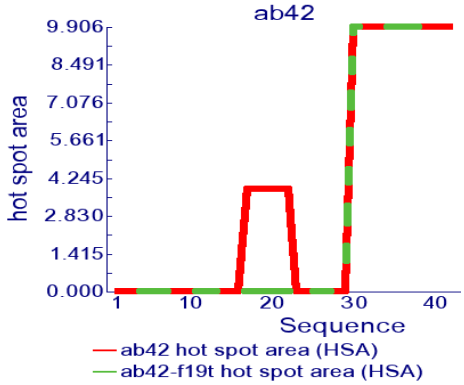

Figure 4: Changes in the hot spot area plot caused by point mutations in amyloidogenic proteins. (Aβ42 wild type (red) and Aβ42 F19T mutant (green).) [Oscar et al., 2007]

3.2.2 AmylPred

This tool uses an algorithm that assorts different methods that have been found or specifically developed to predict features related to the formation of amyloid fibrils (Kimon *et al.,* 2009). The consensus of these methods is defined as a hit if there is an overlap of at least two out of five methods (*Figure 6*) and it is the primary output of the program. AYLPRED shows results of the five methods in text file format as shown in *Figure 5.* Some of the methods include Average Packing Density (G), TANGO and Hexapeptide Conformational Energy (Z). The individual predictions of these methods are maintained on the server for 1 (one) day. Consequently, the tool predicts probable amyloidogenic determinants for a given amino acid sequence of a peptide or protein. AMYLPRED is freely available for academic use at http://biophysics.biol.uoa.gr/AMYLPRED.

```
JOB ID: 1301916245
NAME :  C132F
Time : 14:24:5 , Apr 4 2011

_CONSENSUS PREDICTION_:
6 - 10
26 - 31
60 - 68
79 - 81
106 - 108
148 - 153
162 - 164
183 - 191
206 - 212
255 - 260
277 - 285
302 - 306

_DETAILS_:
 ->* The presence of values in a column indicates a hit. For methods that return scores, the scores
     are used to mark the hit. Otherwise the aminocid sequence of the region is repeated into the
     column. SecStr hits are marked as "a~b" for no reason other than to improve readability.

 Columns - Values
      AA : Residue numbering
     SEQ : Sequence
   graph : Joint prediction histogram showing the number of methods with positive results (hits) at each
           aminoacid residue position.
       G : Sequence regions with Average Packing Density values above threshold (Galzitskaya et al., 2006)
       S : Conformational Switches detected by SecStr (Hamodrakas et al., 2007)
       P : Sequence regions that match the Amyloidogenic Pattern (Lopez de la Paz & Serrano, 2004 )
       T : Sequence regions with Tango scores above threshold (Fernandez-Escamilla et al., 2004)
       Z : Sequence regions with Conformational Energy values below threshold (Zhang et al., 2007)
     CON : Sequence regions predicted as hits by the Consensus method (AmylPred)

    AA SEQ | graph      |   G    |  S  |  P  |   T   |     Z     |  CON
------ --- | --------- | ------ | --- | --- | ----- | --------- | ---
     1  M  |   '       |        |     |     |       |           |
     2  A  |   '       |        |     |     |       |           |
     3  E  |   '       |        |     |     |       |           |
     4  K  |   '       |        |     |     |       |           |
     5  F  | * '       |        |     |     |       | -81.871   |
     6  D  | * *       | 21.632 |     |     |       | -44.669   |  D
     7  C  | * *       | 23.322 |     |     |       | -54.713   |  C
     8  H  | * *       | 22.636 |     |     |       | -54.713   |  H
```

Figure 5. AmyPred results of Four and a half LIM domains protein 1 (C132F)

A

CONSENSUS RESULTS

...based on at least 2 out of 5 successful methods. The '#' symbol marks the hits:

```
                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...   1 MTTCSRQFTSSSSMKGSCGIGGGIGGGSSRISSVLAGGSCRAPSTYGGGLSVSSRFSSGG  60
Prediction       ------------------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...  61 ACGLGGGYGGGFSSSSSFGSGFGGGYGGGLGAGFGGGLGAGFGGGFAGGDGLLVGSEKVT 120
Prediction       ------------------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...  121 MQNLNDRLASYLDKVRALEEANADLEVKIRDWYQRQRPSEIKDYSPYFKTIEDLRNKIIA 180
Prediction       ------------------------------------------------------------####

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...  181 ATIENAQPILQIDNARLAADDFRTKYEHELALRQTVEADVNGLRRVLDELTLARTDLEMQ 240
Prediction       ###---------------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...  241 IEGLKEELAYLRKNHEEEMLALRGQTGGDVNVEMDAAPGVDLSRILNEMRDQYEQMAEKN 300
Prediction       ------------------------------------------------###---------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...  301 RRDAETWFLSKTEELNKEVASNSELVQSSRSEVTELRRVLQGLEIELQSQLSMKASLENS 360
Prediction       -----#####-------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
Wild type...  361 LEETKGRYCMQLSQIQGLIGSVEEQLAQLRCEMEQQSQEYQILLDVKTRLEQEIATYRRL 420
Prediction       -------######--------------------------#######--------------

                 ....,....|....,....|....,....|....,....|....,....|...
Wild type...  421 LEGEDAHLSSQQASGQSYSSREVFTSSSSSSSRQTRPILKEQSSSSFSQGQSS      473
Prediction       ------------------------------------------------
```

B

CONSENSUS RESULTS

...based on at least 2 out of 5 successful methods. The '#' symbol marks the hits:

```
                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...       1 MTTCSRQFTSSSSMKGSCGIGGGIGGGSSRISSVLAGGSCRAPSTYGGGLSVSSRFSSGG  60
Prediction       ------------------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...      61 ACGLGGGYGGGFSSSSSFGSGFGGGYGGGLGAGFGGGLGAGFGGGFAGGDGLLVGSEKVT 120
Prediction       ------------------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...      121 MQNLNDPLASYLDKVRALEEANADLEVKIRDWYQRQRPSEIKDYSPYFKTIEDLRNKIIA 180
Prediction       ------------------------------------------------------------####

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...      181 ATIENAQPILQIDNARLAADDFRTKYEHELALRQTVEADVNGLRRVLDELTLARTDLEMQ 240
Prediction       ###---------------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...      241 IEGLKEELAYLRKNHEEEMLALRGQTGGDVNVEMDAAPGVDLSRILNEMRDQYEQMAEKN 300
Prediction       ------------------------------------------------###---------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...      301 RRDAETWFLSKTEELNKEVASNSELVQSSRSEVTELRRVLQGLEIELQSQLSMKASLENS 360
Prediction       -----#####-------------------------------------------------

                 ....,....|....,....|....,....|....,....|....,....|....,....|....,....|
R127P...      361 LEETKGRYCMQLSQIQGLIGSVEEQLAQLRCEMEQQSQEYQILLDVKTRLEQEIATYRRL 420
Prediction       -------######--------------------------#######--------------

                 ....,....|....,....|....,....|....,....|....,....|...
R127P...      421 LEGEDAHLSSQQASGQSYSSREVFTSSSSSSSRQTRPILKEQSSSSFSQGQSS      473
Prediction       ------------------------------------------------
```

Figure 6. AmylPred results of (a) Wild type sequence of Keratin, type 1 cytoskeletal 16 and (b) Keratin type I cytoskeletal mutant (R127P).

3.2.3 Average Packing Density (G)

G is used to detect both amyloidogenic and disordered regions in a protein sequence. The mean packing density (number of residues within the given distance from the considered residue) enables the prediction of both amyloidogenic and intrinsically disordered regions from protein sequences. Regions with strong expected packing density are believed to be responsible for amyloid formation, while regions with weak expected packing density correspond to disordered regions. G is an important value for the prediction of both intrinsically disordered and amyloidogenic regions of proteins based on the sequence alone.

The calculations of the expected packing density profile are based on a sliding window-averaging technique. For each peptide and protein, in the prediction of amyloidogenic regions the sliding window size is varied from three to nine residues while the sliding window size is 11 (or 41) residues in the case of intrinsically disordered regions prediction. The packing density profile is calculated as follows; first, the expected packing density is determined for each residue then, these numbers are averaged for five residues inside the window and assigned to the central residue of the window. Therefore, the influence of residues along the sequence flanking each window is included in the calculation. The value of the average expected packing density for every position of the polypeptide chain provides the packing density profile. If more than five residues in a row have values over a specified threshold, this region is predicted to be amyloidogenic. On the other hand, any region having more than 11 (or 41) residues with values below a specified threshold is predicted to be intrinsically disordered.

Values above 21.4 obtained from a five-residue long sliding window are considered hits for amyloidogenic regions whilst 20.4 are hits for intrinsically disordered regions. To evaluate the accuracy of, and confidence in, the method of predicting amyloidogenic regions, a database of 67 peptides that are six-residue fibril formers and 91 peptides that are six-residue fibril nonformers were used. They also used the amino acid sequences of 12 disease-related amyloidogenic proteins and peptides to test their method. The sequences were taken from the SWISS-PROT database (http://us.expasy.org/sprot/). True positive and false positive rates were made to obtain the quality of the predictions and to determine thresholds. Receiver operator characteristic (ROC) curves were then made (Galzitskaya OV *et al.,* 2006).

3.2.4 TANGO (T)

TANGO (http://tango.embl.de/) is a statistical mechanics algorithm that was developed to predict protein aggregation. It calculates the tendency of peptides for β aggregation that is highly correlated to the tendency of amyloid fibril formation. TANGO was based on the experimental data that has identified residues within a protein sequence that promote ordered aggregation and amyloid formation.

It is also takes into account physicochemical principles of b-sheet formation, extended by the assumption that the core regions of the aggregate are fully buried. TANGO algorithm is designed to predict cross-beta aggregation in peptides and denatured proteins and consists of a phase-space encompassing the random coil and 4 possible structural states: β-turn, α-helix, β-sheet aggregation and α-helical aggregation.

To predict cross-β aggregating segments of a peptide TANGO simply calculates the partition function of the phase-space. TANGO was benchmarked against 175 peptides of over 20 proteins and was able to predict the sequences experimentally observed to contribute to the aggregation of these proteins. Further TANGO correctly predicts the aggregation propensities of several disease-related mutations in the Alzheimer's b-peptide. Scores above 5.00% proved to be good indicators of aggregation. TANGO is free for academic use but requires user registration (Fernandez-Escamilla *et al.,* 2004).

3.2.5 Hexapeptide Conformational Energy (Z)

This program threads all hexapeptides of a submitted protein onto the microcrystallic structure of NNQQNY (Z. Zhang *et al.,* 2007). The structure of NNQQNY obtained from yeast prion protein , is used together with residue-based statistical potentials to establish an algorithm that will help identify the amyloid fibril-forming segment of proteins.It was based on the fact that , the application of the residue-based satatistical potentials is computationally more efficient than using the atomic-level potentials .

The residue-based statistical potentials has an added advantage of the possiblity of applying it in whole proteome analysis to investigate evolutionary pressure effect or forcast other latent diseases related to amyloid deposits. Alternatively the program can use a set of over 2500 templates produced by small shifts in the structure of NNQQNY. In a consensus method, the version using only the original structure was used, in favour of speed. Interaction energy calculations, which involved the mapping of the expected six-residue peptide onto each of the template structures, were made.

The residue-based statistical potential was used to evaluate the interaction energy score of the central strand in the nine-strand β-sheet with other strands. The lowest energy score obtained from the template structures was then used to assay the fibril-forming propensity of the peptide. Energy values below -27.00 are considered as hits.

3.3 Statistical Analysis

In determination of the quality of the various predictions, six parameters were used. These include: sensitivity, specificity, accuracy, precision, negative predictive value (NPV) and Matthews correlation Coefficient (MCC). A classifier is a mapping from instances to predicted classes. Given a classifier and an instance, there are four possible outcomes: a positive instance classified as positive is counted as a true positive (*tp*), a negative instance classified as positive is counted as a false positive (*fp*), a negative instance classified as negative is counted as a true negative (*tn*), and a positive instance classified as negative counted as false negative (*fn*).

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Specificity} = \frac{tn}{fp + tn}$$

$$\text{Sensitivity} = \frac{tp}{tp + fn}$$

$$\text{NPV} = \frac{tn}{tn + fn}$$

$$\text{MCC} = \frac{tp \times tn - fn \times fp}{\sqrt{(tp + fn)(tp + fp)(tn + fn)(tn + fp)}}$$

Much attention was paid to the MCC because it has an important property of not being affected by the differing proportion of neutral and pathogenic datasets predicted by the different programs. It gives a more balanced assessment of performance than the other performance measures (Baldi *et. al.,* 2000).

3.3.1 Normalization

Normalization was done to prevent imbalanced dataset that comes as a result of unequal number of neural and mutated cases. This will prevent the results of the parameters from being biased.

To normalize the values, *tn* and *fp* were recalculated to tn^2 and fp^2 as shown below.

Positive cases (cases +), P = *tp* +*fn*

Negative cases (cases -), N = *tn* + *fp*

$$tn^2 = \frac{tn \times (tp + fn)}{(tn + fp)}$$

$$= \frac{tn \times P}{N}$$

$$fp^2 = \frac{fp \times (tp + fn)}{(tn + fp)}$$

$$= \frac{fp \times P}{N}$$

4. Results

All of the prediction methods used in this study are freely available on the Internet. It is a laborious task using the main websites to predict large amounts of data. Mainly because most of the websites require the user to manually make the mutations to the wildtype sequences before using the prediction method. PON-P has integrated an algorithm that can introduce many mutations in a sequence at a time. This makes it faster and easier to get results. PON-P also simultaneously submits the input data provided by the user to five aggregation methods solving the problem of going to the main websites for the predictions.

All the 365 missense variations used for this study was downloaded online. AD&FTDMDB and Ovid Medline were the major databases used. Most of 58 proteins used had more than a single variation. These variations are from literature and have been experimentally proven to introduce an increase, a decrease or no effect on the wildtype sequences.

The methods used in this study are Aggrescan, AmylPred consensus, Average Packing Density (G), TANGO and Hexapeptide Conformational Energy (Z). AmylPred consensus results are based on at least 2 of 5 successful methods used by AmylPred. Average Packing Density (G), TANGO and Hexapeptide Conformational Energy (Z) are part of AmylPred results in a text file format.

The results from Aggrescan were not binary because there were cases that involved mutations that caused an increase, a decrease or no effect to the wild type proteins. In that case the results were divided into Aggrescan A^* (It combines the variations that result in an increase and those that result in a decrease in aggregation against the variations that have no effect on the aggregation), Aggrescan B^* (variations that result in an increase in aggregation against the variations that result in a decrease in the aggregation and those that have no effect on the aggregation) and Aggrescan C^* (variations that result in a decrease in aggregation against the variations that result in an increase and those that have no effect on the aggregation).

4.1 Performance of Prediction Methods

Accuracy, Precision, Specificity, Sensitivity, NPV and MCC are the measures used to evaluate the performance of the prediction methods.

Accuracy is a measure of the degree of similarity between the training and test sets. A high accuracy of about 0.9 therefore means a good performance of the prediction method. Sensitivity values show the probability of correctly predicting a positive outcome whilst values of specificity show probability that the positive prediction is correct (Baldi *et. al.,* 2000). Precision (also termed as the

positive predictive value, PPV) gives the proportion of missense variants with a positive test result that actually caused an increase in aggregation. Negative predictive value, NPV is the proportion of missense variants with a negative test result that do not cause an increase in aggregation. Matthews correlation Coefficient, MCC gives values between -1 and +1. A value of +1 indicates perfect classification accuracy whilst -1 indicates total disagreement. Two independent variables result in a correlation coefficient of 0 (Baldi *et. al.,* 2000).

The values for these measures are presented in Table 2 (calculated from non-normalized values) and Table 3 (calculated from normalized values).

The values of the measures were used to draw bar graphs. Measures calculated from normalized values are shown in *figure 7*.

Bar graphs for individual normalized measures; accuracy, precision, specificity, sensitivity, NPV and MCC are shown in *figures 8,9,10,11,12* and *13* respectively.

Table 2. Performance of prediction methods

	AmylPred Consensus	Tango	G	Z	Aggrescan		
					A*	B*	C*
tp	97	29	103	126	310	146	7
fn	37	19	26	36	39	173	23
tn	200	297	203	161	14	33	138
fp	27	16	29	38	2	13	197
cases +	134	48	129	162	349	184	8
cases -	227	313	232	199	16	13	5
Total	361	361	361	361	365	365	365
Accuracy[a]	0.82	0.90	0.85	0.80	0.89	0.49	0.40
Precision[a]	0.78	0.64	0.78	0.77	0.99	0.92	0.03
Specificity[a]	0.88	0.95	0.88	0.81	0.88	0.72	0.41
Sensitivity[a]	0.72	0.60	0.80	0.78	0.89	0.46	0.23
NPV[a]	0.84	0.94	0.89	0.82	0.26	0.16	0.86
MCC[a]	0.62	0.57	0.67	0.59	0.44	0.12	-0.20

[a]Accuracy, precision, specificity, sensitivity, NPV and MCC calculated from non-normalized values

Table 3.. Performance of prediction methods

	AmylPred Consensus	Tango	G	Z	Aggrescan		
					A*	B*	C*
tp	97	29	103	126	310	146	7
fn	37	19	26	36	39	173	23
tn	200	297	203	161	14	33	138
fp	27	16	29	38	2	13	197
tn[b]	118.1	45.5	112.9	131.1	305.4	228.8	12.4
Fp[b]	15.9	2.5	16.1	30.9	43.6	90.2	17.6
cases +	134	48	129	162	349	319	30
cases -	227	313	232	199	16	46	335
Total	361	361	361	361	365	365	365
Accuracy[b]	0.80	0.78	0.84	0.79	0.88	0.59	0.32
Precision[b]	0.86	0.92	0.86	0.80	0.88	0.62	0.28
Specificity[b]	0.88	0.95	0.88	0.81	0.88	0.72	0.41
Sensitivity[b]	0.72	0.60	0.80	0.78	0.89	0.46	0.23
NPV[b]	0.76	0.71	0.81	0.78	0.89	0.57	0.35
MCC[b]	0.61	0.59	0.68	0.59	0.76	0.18	-0.36

[b]Accuracy, precision, specificity, sensitivity, NPV and MCC calculated from normalized values

Aggrescan A* = ((- and +) against '=')

Aggrescan B* = ((+) against (– and '='))

Aggrescan C* = ((-) against (+ and '='))

(-) = Results of mutations that reduce the aggregation propensity of proteins

(+) = Results of mutations that increase the aggregation propensity of proteins

(=) = Results of mutations that have no effect on the aggregation propensity of proteins.

4.2 Evaluation of Aggregation Prediction Methods

The values of Accuracy, Precision, Specificity and NPV changed significantly after normalization of the values. Sensitivity was not affected by normalization. The MCC of the methods did not have any significant change upon normalization except those of Aggrescan A*, B* and C*. MCC values of Aggrescan A* and B* increased sharply from 0.44 to 0.76 and 0.12 to 0.18 respectively. MCC for Aggrescan C* on the other hand was decreased from -0.20 to -0.36.

In this study, only the results from the normalized values were considered. Aggrescan A* performed best in accuracy (0.88), sensitivity (0.89), NPV (0.89) and MCC (0.76). Tango performed best in precision (0.92) and specificity (0.95). Aggrescan C* performed worst in all the measures.

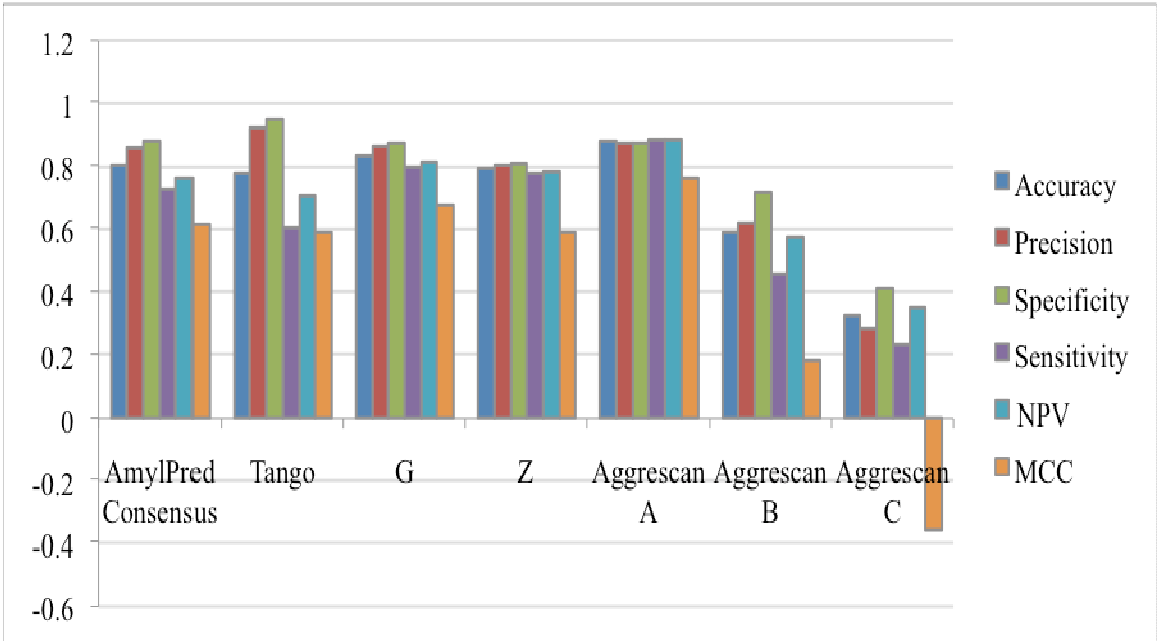

Figure 7: *Bar graph of* accuracy, precision, specificity, sensitivity, NPV and MCC calculated from normalized values

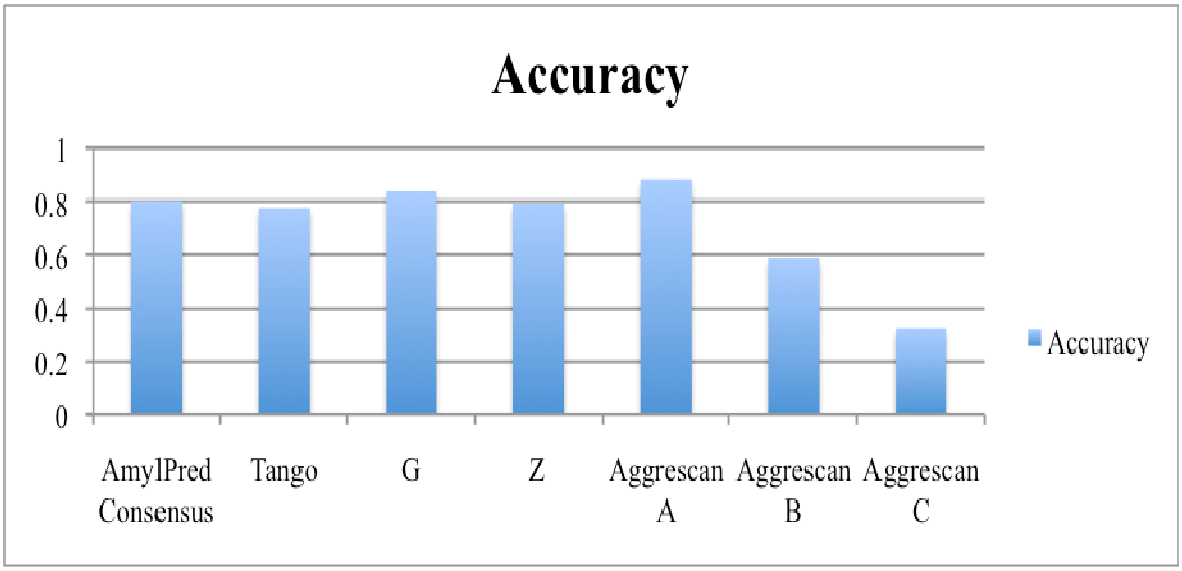

Figure 8: Bar graph of accuracy calculated from normalized values

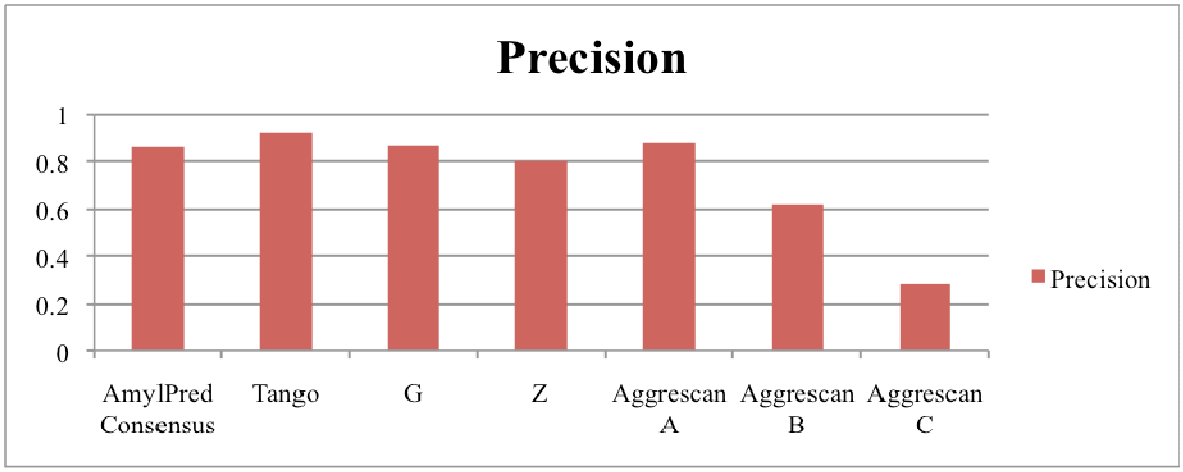

Figure 9: Bar graph of precision calculated from normalized values

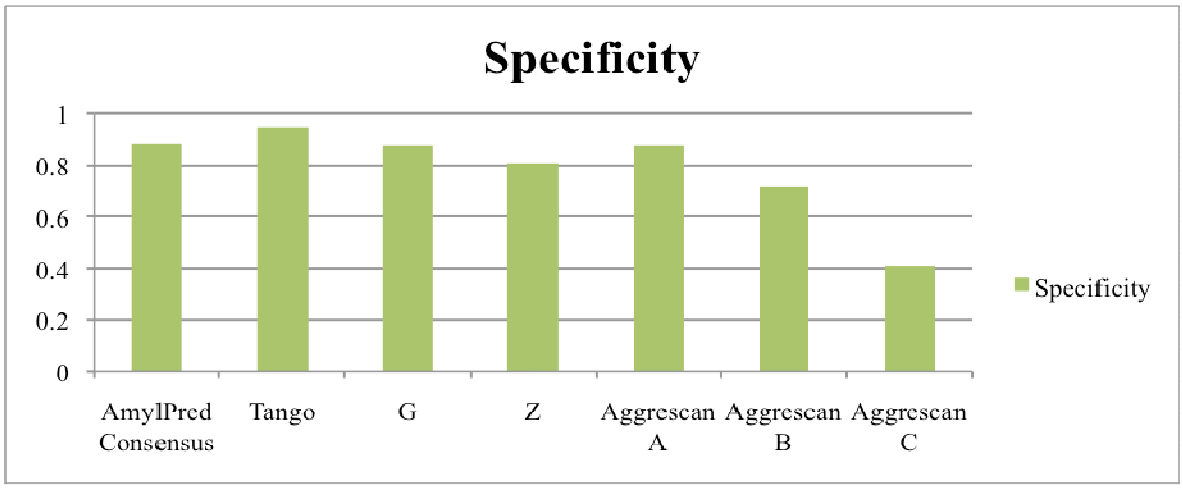

Figure 10: Bar graph of specificity calculated from normalized value

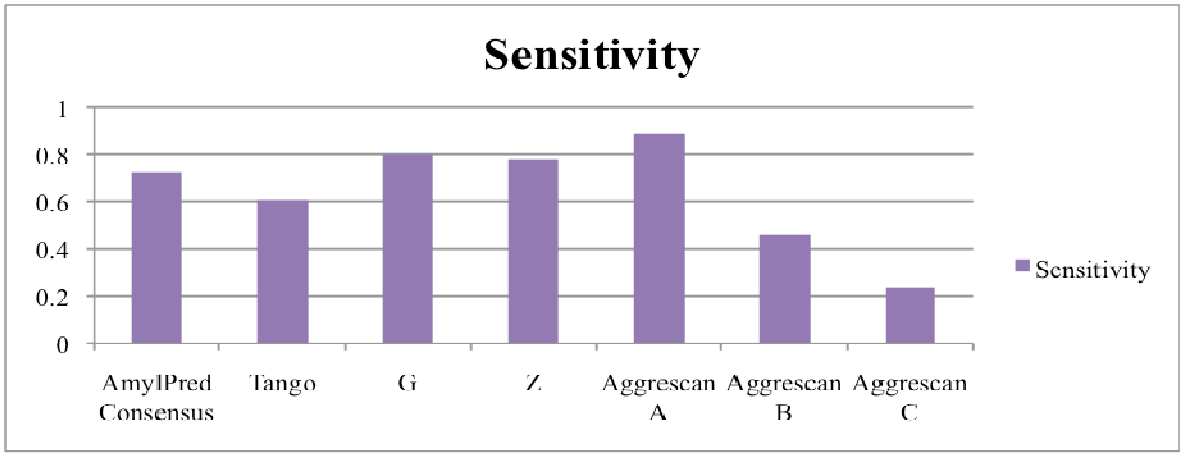

Figure 11: Bar graph of sensitivity calculated from normalized values

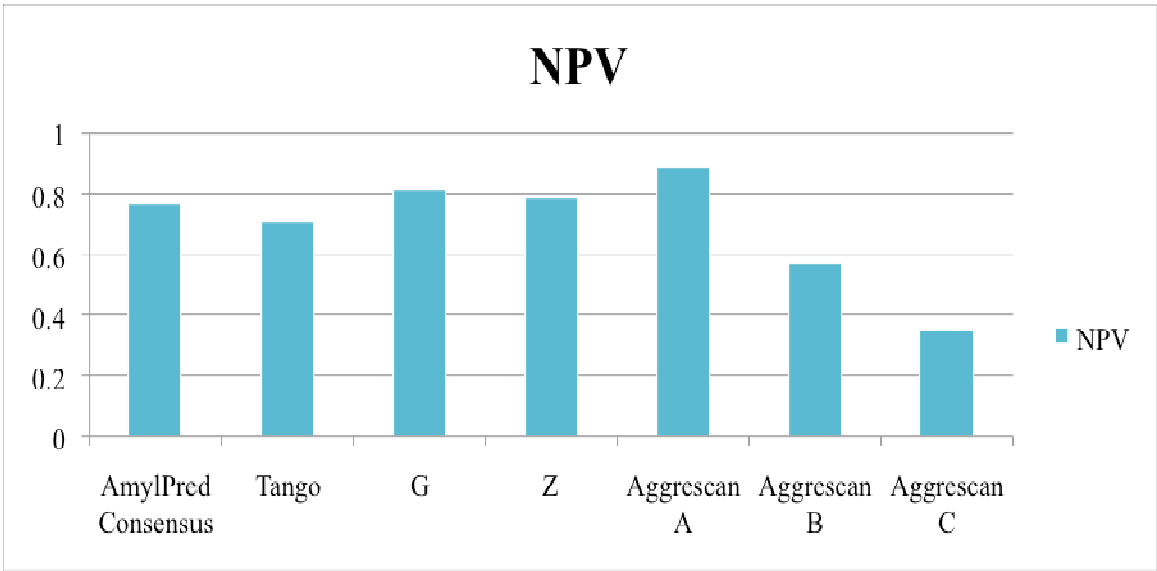

Figure 12: Bar graph of NPV calculated from normalized values

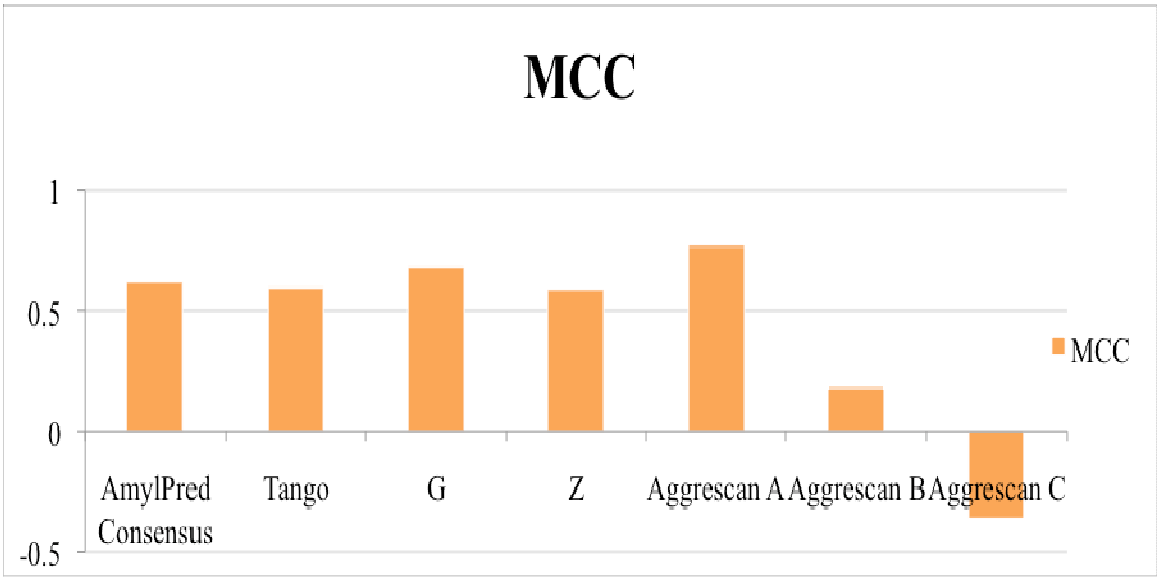

Figure 13: Bar graph of MCC calculated from normalized values

5. Discussion

To be able to know if missense variations are pathogenic or benign, it is important to be able to induce variations into as many proteins as possible and use bioinformatics methods to predict their propensities to aggregate. This will produce enough data for pathogenicity prediction methods. The amount of available identified missense variants has increased rapidly due to the application of high-throughput prediction methods. The number of prediction methods has also been on the rise, most of which are open source.

The need to know the best prediction methods available has therefore increased in recent. All the methods used to predict the aggregation propensity of the proteins in this study are sequence-based.

Even though all the methods use different algorithms, they are all aimed at helping the end user predict aggregation or amyloidogenic prone zones in a protein sequence without going through the laborious task of doing it experimentally.

Experimental results found in literature are used as benchmarks to determine how they correlated with predicted results.

The use of novel data for the study was taken in to much consideration instead of limiting the search to only data that was used in the process of constructing the algorithms of the methods. Due to unequal number of neutral and mutated cases, the numbers of neutral cases were normalized to be equal to the number of mutated cases for each method before calculating the evaluation parameters. This gives a more unbiased result. It can be clearly seen from Tables 2 and 3 that there has been significant changes in the values of the parameters after normalization.

The plots in *figures 8,9,10,11,12* and *13* clearly show the performance of all the methods used in the study. All the methods with the exception of Aggrescan C^* (0.28) had good values for precision with Tango (0.92) being the highest. Tango was also the best in specificity (0.95). Aggrescan A^* performed best in all the other four measures. Aggrescan C^* had the worst values in all the measures.

Considering Aggrescan alone, Aggrescan A^* performed best in all the measures and Aggrescan C^* performing the least in all the measures. Also, considering all the methods except Aggrescan, Average packing density (G) performed best in Accuracy, sensitivity, NPV and MCC.

From the results of the MCC it clearly shows that Aggrescan can be considered the best method even though Aggrescan C^* performed badly. The reason for the low values for Aggrescan C^* can be explained by the fact that it combines the variations that result in an increase and those that have no effect on the aggregation propensities of the proteins against the variations that result in a decrease in the aggregation propensity of the proteins. This therefore shows a very low correlation between the predicted values and the experimental values.

Contrary to previous study that a consensus approach might be better suited for the task of predicting amyloidogenic stretches (Kimon K. F. *et al.,* 2009), the results show that individual prediction methods were better in some cases.

Work by Maria P.C David *et al.,* show that the algorithms used in the prediction methods deal with the prediction of the segments involved or possibly involved in amyloidosis, but do not generate direct predictions on whether a given sequence will be amyloidogenic or not. They therefore proposed artificial intelligence that may be used to complement existing prediction protocols in obtaining direct predictions about the amyloidogenicity of proteins (Maria P.C David *et. al.,* 2010).

6. Conclusion

It can be concluded that Aggrescan performed best in most of the measures applied. Tango was the most precise method and performed best in specificity. It is however not conclusive which method is the most recommended because of environmental factors involved with the different algorithms.

It must also be noted that the data set used in this study is small and contained some of the data used in training the algorithms. Increasing the size of the training set could clearly increase the accuracy, and most of the other measures.

7. References

Aisenbrey Christopher, Borowik Tomasz, Byström Roberth, Bokvist Marcus, Lindström Fredrick, Misiak Hanna, Sani Marc-Antoine , Gröbner Gerhard (2008). How is protein aggregation in amyloidogenic diseases modulated by biological membranes? European Biophysics Journal, Vol. 37 Issue 3, p247.

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. *Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000; 16:412-424.*

Bertram J. (2000). "The molecular biology of cancer". *Mol. Aspects Med.* 21 (6): 167–223.

Branden, C. and Tooze, J. (1999). Introduction to Protein Structure**.** Second Ed., New York, Garland Publishing.

Bryan W. Allen Jr., Charles W. O'Donnell, Matthew Menke, Lenore J. Cowen, Susan Lindquist, Bonnie Berger. "STITCHER: Dynamic assembly of likely amyloid and prion β-structures from secondary structure predictions." Forthcoming in Proteins: Structure, Function, and Bioinformatics, 23 SEP 2011.

Bucciantini Monica, Giulia Calloni, Fabrizio Chiti, Lucia Formigli, Daniele Nosi, Christopher M. Dobson, and Massimo Stefani. (2004). Pre-fibrillar amyloid protein aggregates share common features of cytotoxicity. *J. Biol. Chem.,* 279, 31374–31382.

Bucciantini, M., E. Giannoni, et al. (2002). "Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases." Nature 416(6880): 507-11.

Bucciantini, M., Rigacci, S., Berti, A., Pieri, L., Cecchi, C., Nosi, D., Formigli, L., Chiti, F. and Stefani, M. (2005). Patterns of cell death triggered in two different cell lines by HypF-N pre-fibrillar aggregates. FASEB J. 19, 437-439.

Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, Todd JA, Blundell TL. 2007. Genome bioinformatic analysis of nonsynonymous SNPs. BMC Bioinformatics 8:301.

Chiti F, Dobson CM. (2006). Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75: 333–366.

Chiti, F., M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature. 424:805–808.

Conchillo-Solé, O., N. Sanchez de Groot, F. X. Avilès, J. Vendrell, X. Daura, and S. Ventura. 2007. AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. BMC Bioinformatics. 8:65.

Cruts M, Van Broeckhoven C. Molecular genetics of Alzheimer's disease. Annals of Medicine 30: 560-565, 1998.
Cruts M, Van Broeckhoven C. Presenilin mutations in Alzheimer's disease. *Human Mutation* 11: 183-190, 1998.

Dobson CM. (2003). Protein folding and misfolding. *Nature* 426:887-888.

Fernandez Escamilla, A.M., Rousseau, F., Schymkowitz, J., and Serrano, L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat. Biotech. 22 1302–1306.

Fink A. L. (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Folding Des*. 3:R9–R23.

Galzitskaya OV, Garbuzynskiy SG, Lobanov MV: Prediction of amyloidogenic and disordered regions in protein chains. *PloS Comput Biol* 2006, 2:1639-1648.

Genetic data on *APP*, *PSEN1* and *PSEN2*: Cruts M, Van Broeckhoven C. Molecular genetics of Alzheimer's disease. Annals of Medicine 30: 560-565, 1998 (PubMed ID: 9920359).

Genetic data on *PSEN1* and *PSEN2*: Cruts M, Van Broeckhoven C. Presenilin mutations in Alzheimer's disease. *Human Mutation* 11: 183-190, 1998 (PubMed ID:9521418)

Genetics Home Reference
(http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/neutralmutations)

Gerum C, Schlepckow K, Schwalbe H. 2010. The unfolded state of the murine prion protein and properties of single-point mutants related to human prion diseases. Journal of molecular biology. vol./is. 401/1(7-12), 1089-8638.

Gijselinck I, Van Broeckhoven C, Cruts M. 2008. Granulin mutations associated with frontotemporal lobar degeneration and related disorders: an update. *Human Mutation* 29: 1373-1386.
Hamodrakas SJ, Liappa C, Iconomidou VA. (2007). Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. Int J Biol Macromol. ;41(3):295-300.

Horwich L. and Jonathan S. Weissman. 1997. Deadly Conformations: Protein Misfolding in Prion Disease. Cell, Vol. 89, 499-51

Iconomidou VA, Vriend G, Hamodrakas SJ (2000) Amyloids protect the silkmoth oocyte and embryo. *FEBS Letters* 479, 141-145.

Jiménez JL, Guijarro JI, Orlova E, Zurdo J, Dobson CM, Sunde M, Saibil HR. 1999. Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J.* 18, 815−821.

Kayed R., Head E., Thompson J. L., McIntire T. M., Milton S. C., Cotman C. W. and Glabe C. G.*(2003).* Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis*. Science 300, 486–489.*

Kelly, J.W. (2008). The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* 8, 101−106.

Kimon K Frousios, Vassiliki A Iconomidou, Carolina-Maria Karletidi, Stavros J Hamodrakas, "Amyloidogenic deteminants are usually not buried," BMC Structural Biology 2009, 9:44.

Klunk WE, Jacob RF, Mason RP *(1999)* Quantifying amyloid β-peptide (Aβ) aggregation using the Congo red-Aβ (CR-Aβ) spectrophotometric assay. *Anal Biochem 266: 66–76.*

Lappalainen I, Thusberg J, Shen B, Vihinen M. 2008. Genome wide analysis of pathogenic SH2 domain mutations. Proteins 72:779-792.

Maria Pamela C David, Gisela P Concepcion, Eduardo A Padlan (2010)Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. *BMC Bioinformatics* 2010, 11:79

Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martin IC, Reumers J,Monsellier E, Chiti F. (2007) . Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep. 8: 737-742.

Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martin IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F. 2010. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices Nat Methods. 7(3): 237-42.

Monsellier E, Ramazzotti M, Polverino de Laureto P, Tartaglia GG, Taddei N, et al. (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. Biophys J 93: 4382–4391.

Monsellier E, Ramazzotti M, Taddei N, Chiti F. Aggregation propensity of the human proteome. *PLoS Comput Biol 2008;4:e1000199.*

Parrini C, Taddei N, Ramazzotti M, Degl'innocenti D, Ramponi G, Dobson CM, Chiti F (2005). Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure* **13:** 1143–1151.

Rademakers R, Cruts M, van Broeckhoven C. 2004.The role of tau (MAPT) in frontotemporal dementia and related tauopathies. *Human Mutation* 24: 277-295.

Ross A Christopher, Poirier A Michelle.2004.Protein aggregation and neurodegenerative disease *Nature Medicine* 10, S10–S17

Selkoe DJ (2003): Folding proteins in fatal ways. *Nature* 426: 900-904.

Stefani M, Dobson CM (2003) protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J Mol Med* **81:** 678–699.

Sunde Margaret, Serpell Louise C., Bartlam Mark, Fraser Paul E., Pepys Mark B., Blake Colin C. F. (1997). Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* 273, 729–739.

Tavtigian S, Byrnes G, Goldgar D, Thomas A. 2008a. Classification of rare missense substitutions using risk surfaces, with genetic and molecular epidemiology applications. Hum Mutat 29:1342-1364.

Tavtigian S, Greenblatt M, Lesueur F, Byrnes G. 2008b. In silico analysis of missense mutations using sequence alignment based methods. Hum Mutat 29:1327-1336.

Theuns J, Marjaux E, Vandenbulcke M, Van Laere K, Kumar-Singh S, Bormans G, Brouwers N, Van den Broeck M, Vennekens K, Corsmit E, Cruts M, De Strooper B, Van Broeckhoven C, Vandenberghe R. (2006). Alzheimer dementia caused by a novel mutation located in the APP C-terminal intra-cytosolic fragment. *Human Mutation* 27: 888-896.

Thusberg J, Vihinen M .2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum Mutat*, 30:703-714.

Thusberg J, Vihinen M. 2006. Bioinformatic analysis of protein structure-function relationships: case study of leukocyte elastase (ELA2) missense mutations. Hum Mutat 27:1230-1243.
Thusberg J, Vihinen M. 2007. The structural basis of hyper IgM deficienc-CD40L mutations. Protein Eng. Des Sel 20:133-141.

Thusberg, J., Olatubosun, A. and Vihinen, M. (2011), Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation, 32: 358–368. doi: 10.1002/humu.21445.

Ventura S, Villaverde A .2006. Protein quality in bacterial inclusion bodies. *Trends Biotechnol*, 24(4):179-185.

Ventura, S. & Villaverde, A. (2006). Protein quality in bacterial inclusion bodies. Trends in Biotechnology 24, 179-185.

Walsh, D. M., Klyubin, I., Fadeeva, J. V., Cullen, W. K., Anwyl, R., Wolfe, M. S., Rowan, M. J. and Selkoe, D. J. (2002). Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation invivo.Nature 416, 535-539.

Worth CL, Bickerton GR, Schreyer A, Forman JR, Cheng TM, Lee S, Gong S, Burke DF, Blundell TL. 2007. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. J Bioinform Comput Biol 5:1297-1318.

Zhuqing Zhang, Hao Chen and Luhua La :Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Structural Bioinformatics 2007, Vol. 23 no. 17, pp. 2218–2225.

8. Appendices

Data set (Number of proteins: 58, Number of variations: 365**)**

UniProt ID	Protein Name	Mutant Name	Effect on aggregation ('+' means increase in aggregation '-' means decrease in aggregation '=' means no effect on aggregation)
P49768	Presenilin 1	S290C	+
		M146V	+
		E273A	+
		A231T	+
		L262F	+
		E120D	+
		A246E	+
		I143F	+
		A79V	+
		L113P	+
		V94M	+
		S169P	+
		Y115C	+
		E123K	+
		R269H	+
		T116N	+
		L392P	+
		G209V	+
		M146L	+
		R269G	+
		L250S	+
		T147I	+
		P436Q	+
		L219P	+
		W165G	+
		P264L	+
		L173W	+
		M233L	+
		H163R	+
		P436S	+
		N405S	+
		P117L	+
		N135D	+
		S169L	+
		S390I	+
		A434C	+
		L286V	+
		V261F	+
		E318G	-
		M139I	+
		G209R	+
		E280A	+

		E184D	+
		L235P	+
		E280G	+
		H163Y	+
		R278T	+
		P267S	+
		G378E	+
		V96F	+
		M146I	+
		A231V	+
		L392V	+
		L166R	+
		A426P	+
		M139V	+
		Y115H	+
		W165C	+
		M233T	+
		A285V	+
		I143T	+
		G384A	+
		A260V	+
		L171P	+
		F105L	+
		M139T	+
		C410Y	+
		I213T	+
		A409T	+
		E120K	+
		C263R	+
		V82L	+
		L153V	+
		L219F	+
		M139K	+
P49810	Presenilin 2	M239V	+
		T122P	+
		M239I	+
		N141I	+
		V148I	+
		A91S	+
		L111M	+
		F64L	+
		A45S	+
		C10R	+
		A25T	+
		V20I	+
		S50I	+
		T60A	+
		I84T	+
		T49G	+
		G47E	+
		Y78F	+
		D18E	+
		I107V	+
		F64S	+

		G53E	+
		F64S	+
		L55R	+
		A36P	+
		Y116S	+
		A109S	+
		T49A	+
		V30L	+
		F33V	+
		K35N	+
		Y69H	+
		E42D	+
		I84Q	+
		T59K	+
		L58H	+
		F33I	+
		A45T	+
		E51G	+
		E42G	+
		V28M	+
		L12P	+
		E89K	+
		T49I	+
		E54G	+
		V71A	+
		I73V	+
		I84N	+
		A45D	+
		Y114C	+
		V30M	+
		G47V	+
		S77F	+
		S50R	+
		A36G	+
		R34T	+
		G47A	+
		A120S	+
		S77Y	+
		V122A	+
		S112I	+
		V30G	+
		D38A	+
		A97G	+
		E54K	+
		F33L	+
		I84S	+
		H56R	+
		I107M	+
		V122I	+
		F44S	+
		I68L	+
		V30A	+
		K70N	+
		L58R	+

		S52P	+
		D18G	+
P92624	Amyloid precursor protein	A692G	+
		F690G	+
		A692P	+
		V717G	+
		V717I	+
		E693K	+
		V715M	+
		L723P	+
		E693G	+
		I716V	+
		T714I	+
		V717L	+
		V717F	+
		E693Q	+
P02647	Apolipoprotein A-I	R173P	+
		L60R	+
	Lysosyme	I56T	+
		W64R	+
		D67H	+
	Immunoglobulin-light chain, LEN	L15P	+
		S28F	+
	Immunoglobulin-light chain (Bence Jones) REI	A84T	+
		G57E	+
		G68D	+
		R61N	+
		D82I	+
P56544	Acylphosphatase-1	G15A	-
		G19A	-
		G37A	-
		G45A	-
		G53A	-
		G69A	-
P23202	Protein URE2	R17C	+
O95292	Vesicle-associated membrane protein-associated protein	P56S	+
P17661	Desmin	S13F	+
		R16C	+
		S46F	+
		S46Y	+
		R350P	+
		R454W	+
		E413K	+
		R406W	+
		L345P	+
		L385P	+
P0A334	Voltage-gated potassium channel	T74V	+
		V76I	+
P00441	Superoxide dismutase [Cu-Zn]	G94A	+
		H44R	+
		G86R	+
		G98R	+
P02489	Alpha-crystallin A chain	R49C	+

		R116C	+
		Y118D	-
P14136	Glial fibrillary acidic protein	R239C	+
		R416W	+
P60891	Ribose-phosphate pyrophosphokinase 1	N114S	-
Q9UBF9	Myotilin	S55F	+
P08670	Vimentin	R113C	+
P00918	Carbonic anhydrase 2	H107Y	-
		E237H	+
P37840	Alpha-synuclein	A30P	+
		A53T	+
		E46K	+
		V66P	-
		T72P	-
		T75P	-
		E83A	+
		E126A	=
		S129A	=
Q6PHP7	Crystallin, gamma B	S11R	+
P02545	Prelamin-A/C	R89L	+
		R101P	+
		R166P	+
		R190Q	+
		E203K	+
		I210S	+
		L215P	+
		R482W	=
		R386K	+
		N195K	+
		D192G	+
		L85R	=
		S143F	+
P30131	Carbamoyltransferase hypF	N4T	=
		Q10A	=
		R14K	=
		Q18A	-
		R23K	+
		Q28A	=
		N34A	-
		D38A	=
		N41A	-
		E47A	+
		R49K	=
		D53A	-
		E55A	-
		V59A	+
		C65A	+
		L68A	=
		D72A	-
		E75A	-
		E77A	-
		Q83A	=
		E87A	-

P63261	Actin, cytoplasmic 2	K118M	+
		T278I	+
		P332A	+
		V370A	+
		E241K	+
O43918	Autoimmune regulator	K221A	+
		K222A	+
		K222E	+
		R257A	+
P04156	Major prion protein	V203I	-
		R208H	-
		E196K	+
		F198S	+
P13647	Type II cytoskeletal 5	S181P	+
		I183M	+
		E475G	+
		V186L	+
P13646	Type I cytoskeletal 13	M108T	+
		L115P	+
P02533	Type I cytoskeletal 14	R125H	+
Q16595	Frataxin, mitochondrial	I154F	+
		W155R	+
		D122Y	+
		G130V	+
P48039	Melatonin receptor type 1A	N124L	+
		N124A	+
		N124K	+
Q13148	TAR DNA-binding protein	A315T	+
		G348C	+
		A382T	+
Q9Y487	V-type proton ATPase 116 kDa subunit a isoform 2	P792R	+
		P405L	+
		P87L	+
P58012	Forkhead box protein L2	S58L	+
		I63T	+
		A66V	+
		E69K	+
		S70I	+
		I80T	+
		I84N	+
		F90S	+
		W98G	+
		S101R	+
		I102T	+
		R103C	+
		H104R	=
		L106F	+
		N109K	=
		S217F	=
P55072	Transitional endoplasmic reticulum ATPase	K251A	-
		K524A	-
		T761E	+
		R95G	+
		R155C	+

		R155H	+
		R155P	+
		R191Q	+
		A232E	+
Q5RCS6	Alpha-actinin-4	K255E	+
P07320	Gamma-crystallin D	V76D	+
		H23T	+
Q99574	Neuroserpin	S49P	+
		S52R	+
P12277	Creatine kinase B-type	D54G	+
Q9R0H5	Keratin, type II cytoskeletal 71	A143G	+
		I146F	+
Q86YB8	ERO1-like protein beta	G252S	+
		H254Y	+
O60260	E3 ubiquitin-protein ligase parkin	R275W	+
		C289G	+
		C418R	+
		C441R	+
		R42P	+
		T240R	-
Q14203	Dynactin subunit 1	G59S	+
Q9UJYI	Heat shock protein beta-8	K141E	+
		K141N	+
P02511	Alpha-crystallin B chain	R120G	+
		G154S	+
Q9UJY1	Myosin-reactive immunoglobulin light chain variable region	D82I	+
		R61N	+
P10275	Androgen receptor	K632A	+
		K633A	+
P04637	Cellular tumor antigen p53	R248Q	+
P10636	Microtubule-associated protein tau	R5L	+
O60500	Nephrin	D819V	+
Q01453	Peripheral myelin protein 22	G150D	+
		L16P	+
P35499	Sodium channel protein type 4 subunit alpha	R672G	+
P05067	Amyloid beta A4 protein	D678N	+
Q13642	Four and a half LIM domains protein 1	C101F	-
		C104R	+
		W122S	-
		H123Y	=
		C132F	+
		C209R	-
		C276Y	+
Q99972	Myocilin	E323	+
		G364V	+
		K423E	+
		D380A	+
		P370L	+
O95278	Laforin	T194I	+
		G279S	+
		Y294N	+
P08779	Keratin, type I cytoskeletal 16	R127P	+
		Q122P	+