Dévelopement d'une méthode bio-informatique pour la prédiction des régions amyloidogéniques dans les protéines.

Abdullah Ahmed

Université Montpellier II

BIOLOGIE SANTE

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Ecole Doctorale : Sciences chimique et biologique pour la santé

Discipline : Biologie santé

Présentée et soutenue publiquement

Par

Ahmed Abdullah

Le 2 juillet 2012

Développement d'une méthode bioinformatique pour la prédiction des régions amyloïdogeniques dans les protéines

Development of a bioinformatics based method for the prediction of amyloidogenic regions in proteins

Jury

Dr. Andrea Parmeggiani	Examinateur
Dr. Salvador Ventura Zamora	Rapporteur
Dr. Javier Sancho	Rapporteur
Dr. Andrey Kajava	Directeur de Thèse

Résumé :

La formation d'agrégats protéiques insolubles et fibreux, appelés fibrilles amyloïdes, est impliquée dans une large variété de maladies humaines. Parmi elles, figurent entre autres, le diabète de type II, l'arthrite rhumatoïde et, notamment, les atteintes neurodégénératives débilitantes, telles que les maladies d'Alzheimer, de Parkinson ou encore de Huntington. Actuellement, il n'existe ni traitement, ni diagnostic précoce pour aucune de ces maladies.

De nombreuses études ont montré que la capacité à former des fibrilles amyloïdes est une propriété inhérente à la chaîne polypeptidique. Ce constat a conduit au développement d'un certain nombre d'approches computationnelles permettant de prédire les propriétés amyloïdogéniques à partir de séquences d'amino-acides. Si ces méthodes s'avèrent très performantes vis à vis de courts peptides (~ 6 résidus), leur application à des séquences plus longues correspondant aux peptides et protéines en lien avec les maladies, engendre un nombre trop élevé de faux positifs.

Le principal objectif de cette thèse consiste à développer une meilleure approche bioinformatique, capable de prédire les régions amyloïdogéniques à partir d'une séquence protéique.

Récemment, l'utilisation de nouvelles techniques expérimentales a permis de mieux appréhender la structure des amyloïdes. Il est ainsi apparu que l'élément caractéristique de la majorité des fibrilles amyloïdes impliquées dans les maladies, était constitué d'une structure étagée (β-arcade), résultant de l'empilement de motifs « feuillet β – coude – feuillet β » appelés « β-arches ». Nous avons mis à profit cette particularité structurale pour créer une approche bioinformatique permettant de prédire les régions amyloïdogéniques d'une protéine à partir de l'information contenue dans sa séquence. Les résultats provenant de l'analyse des structures de type β-arcade, connues et modélisées, ont été compilés et traités à l'aide d'un algorithme écrit en langage Java, afin de créer le programme ArchCandy.

L'application de ce programme à une sélection de séquences protéiques et peptidiques, connues pour leur lien avec les maladies, a permis de démontrer qu'il était en mesure de prédire correctement la majorité de ces séquences, de même que les séquences mutées impliquées dans les maladies familiales. Outre la prédiction de régions à haut potentiel amyloïde, ce programme suggère la conformation structurale adoptée par les fibrilles amyloïdes.

Le séquençage de génomes entiers devenant toujours plus abordable, notre méthode offre une perspective de détermination individuelle des profils à risque, vis à vis de maladies neurodégénératives, liées à l'âge ou autres. Elle s'inscrit ainsi pleinement dans l'ère de la médecine personnalisée.

Abstract:

A broad range of human diseases are linked to the formation of insoluble, fibrous, protein aggregates called amyloid fibrils. They include, but are not limited to, type II diabetes, rheumatoid arthritis, and perhaps most importantly, debilitating neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and Huntington's disease. There currently exists no cure, and no means of early diagnosis for any of these diseases.

Numerous studies have shown that the ability to form amyloid fibrils is an inherent property of the polypeptide chain. This has lead to the development of a number of computational approaches to predict amyloidogenicity by amino acid sequences. Although these methods perform well against short peptides (~6 residues), they generate an unsatisfactory high number of false positives when tested against longer sequences of the disease-related peptides and proteins.

The main objective of this thesis was to develop an improved bioinformatics based approach to predict amyloidogenic regions from protein sequence.

Recently new experimental techniques have shed light on the structure of amyloids showing that the core element of a majority of disease-related amyloid fibrils is a columnar structure (β—arcade) produced by stacking of β-strand-loop-β-strand motifs called "β-arches". Using this structural insight, we have created a bioinformatics based approach to predict amyloidogenic regions from protein sequence information. Data from the analysis of the known and modeled β-arcade structures was incorporated into a rule based algorithm implemented in the Java programming language to create the ArchCandy program.

Testing it against a set of protein and peptide sequences known to be related to diseases has shown that it correctly predicts most of these sequences and a number of mutated sequences related to the familial diseases. In addition to the prediction of regions with high amyloidogenic potential, a structural arrangement of the amyloid fibril is also suggested for each prediction. As whole genome sequencing becomes cheaper, our method provides opportunity to create individual risk profiles for the neurodegenerative, age-related and other diseases ushering in an era of personalized medicine.

Table of Contents

1. Introduction

1.1 What are Amyloids and Why are they Important?

"Amyloid" is primarily used to describe extracellular, fibrous, proteinaceous deposits in organs and tissues. However, they have also been shown to form inside cells and *in vitro*. They are formed by the self assembly of normally soluble proteins into insoluble fibrils resistant to degradation. Scientific interest in them is primarily motivated by the fact that they are involved in several diseases. This section is a general introduction to what is known about amyloids, and their roles in living organisms.

The term amyloid was originally coined by Matthias Jakob Schleiden in 1838 to describe the starchy component of plants. It was derived from the Latin word for starch, *amylus*, which is in turn derived from the Greek *amulos* meaning not ground at a mill. In 1854, amyloid was used for the first time in a human context by Rudolph Virchow to describe extracellular deposits found in several human organs (cerebral corpora, liver, and spleen) (Virchow 1854). The name came from his understanding that the deposits were composed of starch, as they stained pale blue with iodine, and then violet upon treatment with sulphuric acid. In 1859, Friedreich and Kekule revealed, via measurements of nitrogen content, the presence of protein and the absence of carbohydrates in amyloid deposits (Friedreich 1859). However, the incorrect classification stuck. Even though they are highly proteinaceous in nature, this class of extracellular deposits continues to be described as amyloids to this day.

Amyloid fibrils became easier to identify in 1922 when it was shown by Bennhold that Congo red dye binds to them and produces apple green birefringence (Figure 1A) (Bennhold 1922). Subsequently, it was shown that the dye thioflavin T (ThT) also binds to amyloids. Using a combination of histopathological techniques by the 1950's it was established that amyloid deposits occur in the heart, intestines, tongue, liver, lungs, spleen, brain, adrenal glands, and skeletal muscles (Symmers 1956). Since then details of amyloid fibrils were further elucidated by the arrival of new techniques. Transmission electron micrographs confirmed that amyloids are composed of fibrils (Figure 1B) (Cohen and Calkins 1959; Sunde and Blake 1997). X-ray diffraction

analyses revealed the high β-sheet content of the fibrils and the "cross-β" pattern (Figure 1C and 1D) (Astbury *et al.* 1935; Eanes and Glenner 1968; Bonar *et al.* 1969; Sunde *et al.* 1997).

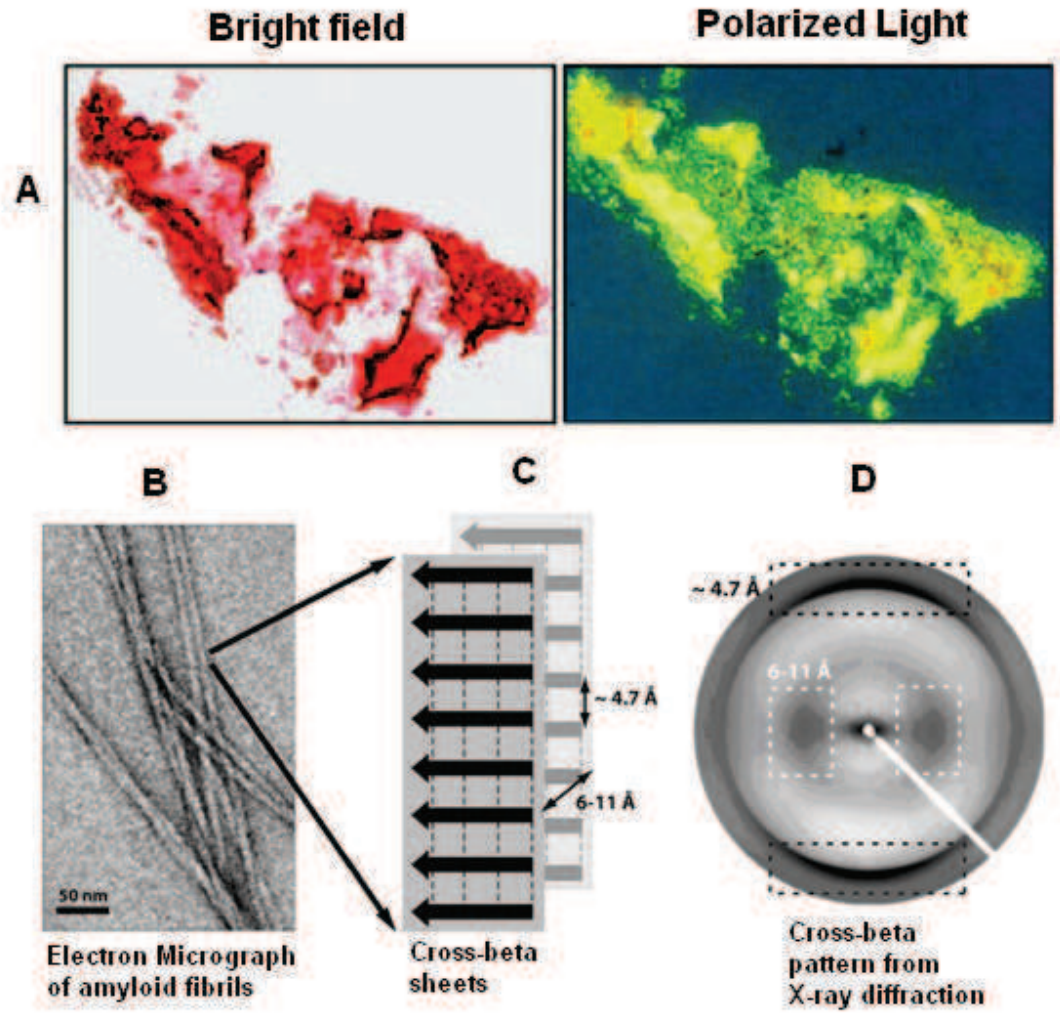


Figure 1. *A: Staining with Congo red and the apple green birefringence under cross polarized light side by side. B: An electron micrograph of amyloid fibril. C: The cross β pattern. The direction of the β-strands (shown by arrows) is perpendicular to the fibril axis. It also shows that the distance between sheets is 6-11 Å and the distance between each strand is ~4.7 Å. D: Two major reflections found during X-ray diffraction. They represent the distances between β-sheets and β-strands. Figure adopted from (Greenwald and Riek 2010)*

This pattern was first observed in the analysis for silk from the egg stalk of the lacewing, Chrysopa (Geddes *et al.* 1968). The name is derived from the fact that individual β-strands lie perpendicular to the fibril axis and the direction of the β-sheet is perpendicular to it, forming a cross. It was also seen that the β-sheets in the core of the fibril form hydrogen bonds between β-sheets parallel to the direction of the fibril axis. Two major reflections occur in the diffraction pattern at ~4.7 Å and 6-11 Å. They represent the hydrogen bonding distance between the β-strands, and the distance of side-chain packing between sheets respectively.

Until this point in history most knowledge on amyloids was mainly generated by scientific curiosity in these strange structures or by fortuitous accidents. This drastically changed with the realization that amyloid fibrils are involved in disease. In the 1980's it was discovered that the main component of amyloid plaques formed in Alzheimer's disease was the Amyloid-β peptide (Glenner and Wong 1984). It is now known that a variety of human diseases are associated with amyloid fibril formation. They include, but are not limited to, type II diabetes, rheumatoid arthritis, and perhaps most importantly, debilitating neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and Huntington's disease. Table 1 shows a list of human diseases associated with the formation of extracellular amyloid deposits or intracellular inclusions (Chiti and Dobson 2006). Furthermore, amino acid sequence analysis of ex-vivo fibrils showed that each amyloid disorder was associated with a specific protein or peptide (Glenner *et al.* 1971). In 1982 the "prion" hypothesis was put forth by Prusiner to explain the infectious cycle of a fatal, degenerative disease that affects the nervous systems of sheep and goats called Scrapie (Prusiner 1982). It stated that the infectious agent of the disease was not another organism but in fact a misfolded protein particle. Which when transmitted to a healthy organism can induce amyloidogenesis in the correctly folded form of the protein. These fibrils can then induct other copies of the protein into the prionic form. A few years later the Bovine Spongiform Encephalopathy (BSE) epidemic in cattle in the UK refocused attention on prions. Efforts to understand prions were substantially increased by the emergence of Creutzfeldt-Jacob disease: a fatal, prion disease whose transmission to humans was linked to the BSE agent (Kretzschmar and Tatzelt 2013).

Table 1 Human diseases associated with formation of extracellular amyloid deposits or intracellular inclusions with amyloid-like characteristics

Disease	Aggregating protein or peptide	Number of residues [a]	Native structure of protein or peptide [b]
Neurodegenerative diseases			
Alzheimer's disease [c]	Amyloid β peptide	40 or 42 [f]	Natively unfolded
Spongiform encephalopathies [c,e]	Prion protein or fragments thereof	253	Natively unfolded (residues 1–120) and α-helical (residues 121–230)
Parkinson's disease [c]	α-Synuclein	140	Natively unfolded
Dementia with Lewy bodies [c]	α-Synuclein	140	Natively unfolded
Frontotemporal dementia with Parkinsonism [c]	Tau	352–441 [f]	Natively unfolded
Amyotrophic lateral sclerosis [c]	Superoxide dismutase 1	153	All-β, Ig like
Huntington's disease [d]	Huntingtin with polyQ expansion	3144 [g]	Largely natively unfolded
Spinocerebellar ataxias [d]	Ataxins with polyQ expansion	816 [g,h]	All-β, AXH domain (residues 562–694); the rest are unknown
Spinocerebellar ataxia 17 [d]	TATA box-binding protein with polyQ expansion	339 [g]	α+β, TBP like (residues 159–339); unknown (residues 1–158)
Spinal and bulbar muscular atrophy [d]	Androgen receptor with polyQ expansion	919 [g]	All-α, nuclear receptor ligand-binding domain (residues 669–919); the rest are unknown
Hereditary dentatorubral-pallidoluysian atrophy [d]	Atrophin-1 with polyQ expansion	1185 [g]	Unknown
Familial British dementia [d]	ABri	23	Natively unfolded
Familial Danish dementia [d]	ADan	23	Natively unfolded
Nonneuropathic systemic amyloidoses			
AL amyloidosis [c]	Immunoglobulin light chains or fragments	~90 [f]	All-β, Ig like
AA amyloidosis [c]	Fragments of serum amyloid A protein	76–104 [f]	All-α, unknown fold
Familial Mediterranean fever [c]	Fragments of serum amyloid A protein	76–104 [f]	All-α, unknown fold
Senile systemic amyloidosis [c]	Wild-type transthyretin	127	All-β, prealbumin like
Familial amyloidotic polyneuropathy [d]	Mutants of transthyretin	127	All-β, prealbumin like
Hemodialysis-related amyloidosis [c]	β2-microglobulin	99	All-β, Ig like
ApoAI amyloidosis [d]	N-terminal fragments of apolipoprotein AI	80–93 [f]	Natively unfolded
ApoAII amyloidosis [d]	N-terminal fragment of apolipoprotein AII	98 [i]	Unknown
ApoAIV amyloidosis [c]	N-terminal fragment of apolipoprotein AIV	~70	Unknown
Finnish hereditary amyloidosis [d]	Fragments of gelsolin mutants	71	Natively unfolded
Lysozyme amyloidosis [d]	Mutants of lysozyme	130	α+β, lysozyme fold
Fibrinogen amyloidosis [d]	Variants of fibrinogen α-chain	27–81 [f]	Unknown
Icelandic hereditary cerebral amyloid angiopathy [d]	Mutant of cystatin C	120	α+β, cystatin like
Nonneuropathic localized diseases			
Type II diabetes [c]	Amylin, also called islet amyloid polypeptide (IAPP)	37	Natively unfolded

……../……..

Disease	Aggregating protein or peptide	Number of residues [a]	Native structure of protein or peptide [b]
Medullary carcinoma of the thyroid [c]	Calcitonin	32	Natively unfolded
Atrial amyloidosis [c]	Atrial natriuretic factor	28	Natively unfolded
Hereditary cerebral haemorrhage with amyloidosis [d]	Mutants of amyloid β peptide	40 or 42 [f]	Natively unfolded
Pituitary prolactinoma	Prolactin	199	All-α, 4-helical cytokines
Injection-localized amyloidosis [c]	Insulin	21 + 30 [j]	All-α, insulin like
Aortic medial amyloidosis [c]	Medin	50 [k]	Unknown
Hereditary lattice corneal dystrophy [d]	Mainly C-terminal fragments of kerato-epithelin	50–200 [f]	Unknown
Corneal amylodosis associated with trichiasis [c]	Lactoferrin	692	α+β, periplasmic-binding protein like II
Cataract [c]	γ-Crystallins	Variable	All-β, γ-crystallin like
Calcifying epithelial odontogenic tumors [c]	Unknown	~46	Unknown
Pulmonary alveolar proteinosis [d]	Lung surfactant protein C	35	Unknown
Inclusion-body myositis [c]	Amyloid β peptide	40 or 42 [f]	Natively unfolded
Cutaneous lichen amyloidosis [c]	Keratins	Variable	Unknown

[a] Data refer to the number of residues of the processed polypeptide chains that deposit into aggregates, not of the precursor proteins.

[b] According to Structural Classification Of Proteins (SCOP), these are the structural class and fold of the native states of the processed peptides or proteins that deposit into aggregates prior to aggregation.

[c] Predominantly sporadic, although in some cases hereditary forms associated with specific mutations are well documented.

[d] Predominantly hereditary, although in some cases sporadic forms are documented.

[e] Five percent of the cases are transmitted (e.g., iatrogenic).

[f] Fragments of various lengths are generated and have been reported to be present in ex vivo fibrils.

[g] Lengths shown refer to the normal sequences with nonpathogenic traits of polyQ.

[h] Length shown is for ataxin-1.

[i] The pathogenic mutation converts the stop codon into a Gly codon, extending the 77-residue protein by 21 additional residues.

[j] Human insulin consists of two chains (A and B, with 21 and 30 residues, respectively) covalently linked by disulfide bridges.

[k] Medin is the 245–294 fragment of human lactadherin.

Adopted from (Chiti and Dobson 2006).

Alzheimer's disease is just one of the debilitating neurological diseases linked to fibril deposits, however, the problems associated with it are representative of the issues related to other amyloid diseases. It is the sixth leading cause of death in the United States after heart disease, cancer, chronic lower respiratory diseases, stroke, and unintentional accidents (Deaths: Final Data for 2010. NVSR Volume 61, Number 04. Accessible at http://www.cdc.gov/nchs/products/nvsr.htm). Nonetheless, it is the only top ten leading cause of death in America that does not have a cure, a means of prediction of predisposition to it, or even a way to stop the progression of disease. Deaths from Alzheimer's increased 68 percent between 2000 and 2010, while deaths from other major diseases, including the number one cause of death (heart disease), decreased (figure 2) (Alzheimer's Disease Facts and Figures Report 2013, Alzheimer's Association. http://www.alz.org/alzheimers_disease_facts_and_figures.asp)

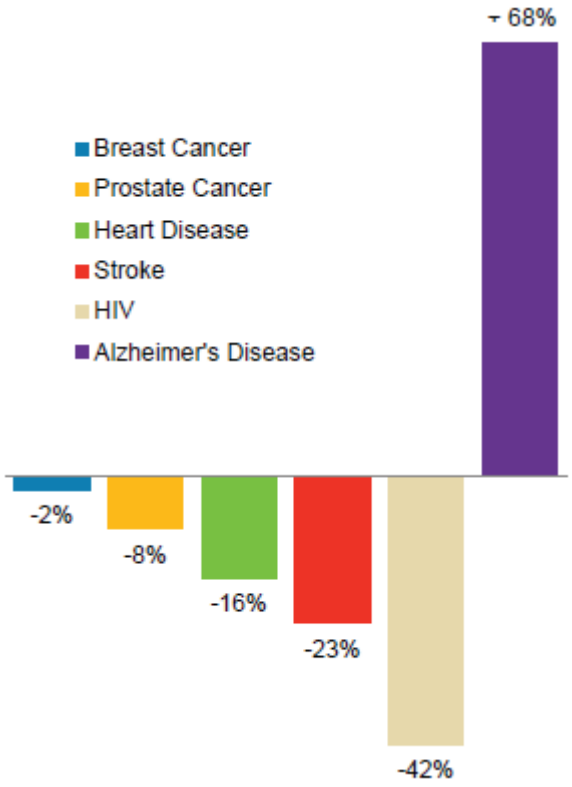


Figure 2. *Change in the number of deaths between 2000 and 2010. Adopted from Alzheimer's Disease Facts and Figures Report 2013* (http://www.alz.org/alzheimers_disease_facts_and_figures.asp).

Another issue associated with Alzheimer's disease is an age related increase in its prevalence. While, approximately 1 percent among those 65 to 69 years of age have the disease, this increases to 40 to 50 percent among persons 95 years of age and over (Hy and Keller 2000). With better standards of living and improved health care the median age, and hence, the population above 65 years of age will rise. By 2025, the number of people age 65 and older with Alzheimer's disease is estimated to reach 7.1 million—a 40 percent increase from the 5 million aged 65 and older currently affected. By 2050, the number of people age 65 and older with Alzheimer's disease may nearly triple, from 5 million to a projected 13.8 million, barring the development of medical breakthroughs to prevent, slow or stop the disease (Hebert *et al.* 2003).

Currently research into amyloids is mainly motivated by the fact that there currently exists no cure, no means of halting fibril formation or preventing it, and no methods for the early diagnosis for any of these diseases.

However, it is important to note that amyloid fibril formation is not always associated with the improper processing or folding of amino acid sequences. The multitude of divergent paths taken by evolution has also resulted in the fascinating development of biologically functional fibrils. Several types of functional amyloids are known to exist in bacteria and fungi. Curli proteins in *Escherichia coli* are involved in the colonization of inert surfaces by biofilm formation and binding to host proteins *(Olsen et al. 1993; Vidal et al. 1998; Chapman et al. 2002)*. Hydrophobins in fungi participate in the formation of hydrophobic aerial structures like aerial hyphae, spores and fruiting bodies (Wessels 1997; Wosten and de Vocht 2000; Wosten and Willey 2000). Chaplins in *Streptomyces coelicolor* form amyloid fibrils that lower the surface tension of water to allow aerial growth. They also cover these structures, making them hydrophobic (Claessen *et al.* 2003). Bacteriocins are antibacterial proteins that act by forming ion channels in membranes, degrading DNA, blocking protein translation, or inhibiting peptidoglycan synthesis (Riley 1998). Microcin E492 in *Klebsiella* pneumoniae is harmless in the amyloid form but has antibacterial activity otherwise (de Lorenzo 1984; Bieler *et al.* 2005).

It has also been suggested that the prion proteins Sup35 and Ure2p from *Saccharomyces cerevisiae* also have functional roles (True and Lindquist 2000). However, the low occurrence of the fibrillated forms of both proteins suggests that the highly specific conditions required for this state to be beneficial occur rarely. Non-fibrillated Sup35 is involved in the termination of mRNA translation. It loses this ability upon fibril formation, however, this allows read through of stop codons leading different phenotypes (True and Lindquist 2000; Marcelino-Cruz *et al.* 2011) . Amyloid fibril formation of Ure2p destroys its ability to sequester the transcription factor Gln3p, resulting in the activation of genes involved in uptake of poor nitrogen sources (True and Lindquist 2000; Chien *et al.* 2004).

Melanin is one of nature's chemical defences against pathogens, small toxic molecules, and UV radiation (Hearing 2000). Recently, it was discovered that functional amyloids participate in the formation of melanin from tyrosine (Fowler *et al.* 2006). The protein Pmel17 acts as a template to position the intermediates of this pathway and accelerates their covalent polymerization into melanin. This also has the beneficial side-effect of sequestering the reactive intermediates (Berson *et al.* 2001; Berson *et al.* 2003).

Amyloids also seem to be involved in the formation of long term memories (Si *et al.* 2003). Although the exact mechanism is not known, it is believed that memory formation requires changes in neuronal synapses, perhaps by protein regulation. The cytoplasmic polyadenylation element binding (CPEB) protein is considered the leading candidate for synaptic translation regulation. It has been shown that it is necessary for long term synaptic changes in *Aplysia* and that it forms amyloid fibrils endogenously in yeast, and exogenously in sensory neurons (Si *et al.* 2010). It has been proposed that the fibrillated form is the active state, and that it provides a long lasting change after a signalling event. It has also been suggested that an increase in the amount of fibrillated protein may act as a means of strengthening the memory after repetitive stimulations of the synapse (Greenwald and Riek 2010).

Amyloids may also have functional roles in humans as a storage mechanism (Maji *et al.* 2009). Some secretory cells can store proteins and peptides for extended periods of time in a highly concentrated form inside membrane enclosed cores called "secretory granules" (Kelly 1985). To test if they were stored as fibrils a study was conducted on 42 randomly selected hormones at pH 5.5. It revealed that in the presence of an aggregation promoting agent (heparin), 31 of the hormones tested are able to form fibrils (Maji *et al.* 2009).

Several benefits have been proposed for this storage mechanism. Firstly, amyloid fibrils are highly sequence specific. Once amyloidogenesis is initiated, further aggregation is self selective. This means that the amyloid itself is able to recruit more proteins. Furthermore, the fibril cores are composed of one hormone only (Greenwald and Riek 2010). The amyloid core also provides the densest packing possible (Nelson *et al.* 2005). Amyloids are believed to have a natural ability to bind to membranes (Sparr *et al.* 2004; Gellermann *et al.* 2005). It is possible that membrane formation around the

hormone fibrils is spontaneous (Greenwald and Riek 2010). Finally, each hormone can have its own disassociation rate which can be controlled by pH, ionic concentration, and/or extracellular chaperons (Greenwald and Riek 2010). Production of the proteins involved in these processes is generally very tightly regulated. These endogenous proteins often originally occur in folded-non amyloidogenic states until required. The transition into an amyloid fibril occurs under tightly controlled conditions. This suggests that fibril formation can have beneficial roles, but only when fibril formation is carefully supervised.

Another reason why amyloids are of interest is due to their role in the production of recombinant proteins. Proteins can aggregate inside cells to produce dense protein deposits called inclusion bodies (IBs) (Kopito 2000). This process occurs more often when large amounts of foreign proteins are produced inside the cells (Marston 1986). IBs were traditionally thought to be disordered aggregates. However, it was recently shown that they are formed by a reaction mechanism that is very similar to that of amyloid formation. Furthermore, like amyloids, they are "seed" aggregation of soluble proteins in a nucleation dependant fashion. This leads to a very interesting situation where understanding amyloidogenesis may lead to means of producing recombinant proteins more efficiently, but the IBs phenomenon itself may also be used as a model for understanding fibril formation (Carrio *et al.* 2005)

In the last two decades the biological roles of amyloids, both in disease and otherwise, have become increasingly clear, and considerable effort has been made to understand their structures, mechanisms of formation, and functions. However, there is a dearth of knowledge on the subject, and this can largely be attributed to several properties of amyloids that make them difficult to study. They are large (mega-dalton) structures with variable lengths and ultra-structural appearances making detailed understanding of the complete structure a lengthy task (Toyama and Weissman 2011). Moreover, their insolubility makes the application of methods traditionally used to elucidate structure, for example solution nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography, impossible. However, progress has been made in this domain with the use of new techniques such as cryoelectron microscopy, scanning transmission electron microscopy, mass measurements, electron paramagnetic resonance, solid state NMR, and the application of existing techniques in innovative ways (Benzinger *et al.* 1998;

Sharma *et al.* 2005; Margittai and Langen 2008; Sachse *et al.* 2008; Goldsbury *et al.* 2011). Introduction section 1.4 provides a more detailed look into the information currently available regarding the three dimensional structure of amyloid fibrils, and how it can be used for their prediction.

It is important to understand that amino acid chains and fibrils are merely the starting and end points. A peptide chain can pass through several intermediate states before forming a fibril (Caughey and Lansbury 2003). There are also off-pathway aggregates that do not fibrillate. The exact aggregation pathway is determined via a combination of the composition of the amino acid sequence, modifications made to it, and the environment within which it is found (Lotz and Legleiter 2013). Several types of intermediates have been observed (Fandrich 2012): Members of the largest class of intermediates are collectively known as oligomers. They do have a specific overall shape, but are generally referred to as spherical (Barghorn *et al.* 2005; Broersen *et al.* 2010). Little is known about them because they are generally kinetic intermediates in the amyloidogenesis pathway and only occur transiently. What is known about them is through equilibrium intermediates which represent balance between the folding and unfolding of the amino acid chain, and can be maintained by keeping environmental conditions steady (Lotz and Legleiter 2013). It is generally believed that oligomers are the toxic component in disease (Stefani and Dobson 2003). There are three proposed methods for their action: they may co-localize with, and sequester housekeeping proteins (Lotz and Legleiter 2013), hence preventing them from carrying out their functions, they may interfere with the cells protein quality control and clearance mechanisms (Bence *et al.* 2001), or they may interact with, and compromise the integrity of cell membranes (Lashuel and Lansbury 2006). It has been shown that cellular models of Huntington's and Parkinson's disease have reduced pathology in the presence of compounds that promote fibrillation. This suggests that their conversion to insoluble, biologically inert fibrils is a mechanism of sequestration and detoxification (Bodner *et al.* 2006). However, it is important to note that amyloid fibrils have significant structural rigidity, and may be able to cause impairments to the tissues where they are deposited. For example, amyloid β fibrils depositing in cerebral blood vessels may weaken them, leading to haemorrhages and stroke (Lotz and Legleiter 2013). Currently, it has not been definitively determined whether oligomers or fibrils are the toxic agents.

Other types of intermediates are closer to mature fibrils. One type is elongated, linear, and high in β-structure; but generally shorter than and lacking in the periodic symmetry of mature fibrils. They also often have weaker binding to CR and ThT (Fandrich 2012). In other cases several intermediates combine to form the fibril (Kajava *et al.* 2010). Annular aggregates do not form fibrils. They have a ring-like shape that encloses a central water filled channel. Not much is known about their structure but they seem similar to pore-forming toxins, and it has been suggested that they may also be toxic because of their ability to perturb the cell membrane (Lashuel *et al.* 2002).

At the end of this introduction to amyloids it is interesting to speculate, in evolutionary terms, where and how amyloids came to be. The conventional view of the evolution of proteins is that evolutionary pressure lead to the development of proteins either with greater efficiency (for example maximum catalytic activity) or new function. However, it has been shown that this is not the only motivation for their evolution. It is in fact influenced by a variety of factors such as the genomic position of the encoding genes, their expression patterns, their position in biological networks and possibly their robustness to mistranslation (Pal *et al.* 2006). It is increasingly becoming accepted that amyloidogenesis is an inherent property of amino acid sequences (Iconomidou and Hamodrakas 2008), and that fibril formation is generally detrimental to the organism. Amyloids can sometimes have functional roles but it should be noted that these proteins are very tightly regulated to prevent uncontrolled fibril formation. This suggests that preventing fibril formation was also a potent force in the evolution of proteins (Dobson 1999). It has also been hypothesised that several proteins have been found to fold very quickly are doing so not only to become functional very quickly, but also to minimize the chances of going towards the competing intermolecular processes of aggregation (Dobson 1999).

Extending this concept further, it has been suggested that amyloids may have been the original conformation of proteins. It has been hypothesised that the first pre-biotic amino acid sequences were amyloidogenic in nature, and were responsible for recruiting membranes and nucleic acids via their ability to bind to repetitive sequences (Greenwald and Riek 2010). According to this hypothesis the ability to form globular domains was not how proteins originally acted, and was in fact acquired through

evolution. Bioinformatics analysis has corroborated this theory. It has been shown that organism complexity inversely correlates with proteomic aggregation propensity (Tartaglia *et al.* 2005).

1.2. Predicting Amyloids from Sequence Data

Progress towards finding a cure for amyloid disease is hindered by the fact that the precise mechanisms of amyloid fibril formation are not known, and all their structural details have not yet been revealed. Bioinformatics tools present an interesting avenue to address these issues. The objective of this research is to develop an easy usable program capable of accurately predicting the potential of amino acid sequences to form fibrils under physiological conditions. Advancement in this direction has the potential to predict individual specific predisposition to amyloid diseases from their genomic data. It has applications in the development of self-assembling nanotechnologies, and drugs that target specific amyloid forming regions in proteins.

Here the approaches and the programs that have been developed to predict the ability of amino acid sequence to form amyloid fibrils based on sequence information are discussed. It should be noted that this is not an exhaustive list. Only the most popular, most diverse in terms of basic principles, and those that can be downloaded or used via web servers are described. Table 2 below shows them.

There are five major approaches to predicting amyloid fibrils. Some methods use only one others use a combination of several approaches:

- Calculation of individual amino acid aggregation propensities.
- Evaluation of properties of β-structural conformation.
- Assessment of the pairwise side-chain to side-chain interactions within β-sheets.
- Methods inspired by the understanding of the amyloid structures of short peptides.
- Estimations of the probability of structured proteins to become partially unfolded.

Table 2. Methods to predict amyloids described here and available online.

Name	Basic approach	Server/Website
AGGRE-SCAN	Composition of amino acids	http://bioinf.uab.es/aggrescan/
Fold-Amyloid	Composition of amino acids	http://bioinfo.protres.ru/fold-amyloid/oga.cgi
Zyggregator	Properties of β-structural conformation	http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php
TANGO	Properties of β-structural conformation	http://tango.crg.es/
PASTA	Pairwise interactions within the β-sheets	http://protein.bio.unipd.it/pasta/
BetaScan	Pairwise interactions within the β-sheets	http://groups.csail.mit.edu/cb/betascan/betascan.html
3D Profile method (ZipperDB)	Amyloid-like structures of short peptides	http://services.mbi.ucla.edu/zipperdb/submit
Waltz	Amyloid-like structures of short peptides	http://waltz.switchlab.org/
NetCSSP	Conformational switches	http://cssp2.sookmyung.ac.kr/index.html
AmylPred	Conformational switches	http://biophysics.biol.uoa.gr/AMYLPRED/

1.2.1 Calculation of individual amino acid aggregation propensities

The ability to form amyloid fibrils is sequence composition dependant. It has been shown that mutations causing simple physico-chemical changes such as hydrophobicity, secondary structure propensity and charge can affect the ability and the rate of fibril formation (Chiti *et al.* 2003). Several approaches have been developed to determine the individual effects of each type of mutation on a proteins ability to fibrillate (DuBay *et al.* 2004; Rojas Quijano *et al.* 2006; Conchillo-Sole *et al.* 2007; Garbuzynskiy *et al.* 2010). These properties are often represented as an amino acid aggregation propensity scale, where a numerical value is assigned to each of the 20 natural amino acids corresponding to their potential to make a sequence more or less likely to undergo amyloid formation. This scale is then exploited by algorithms in various ways to determine the aggregation potential of a given sequence. Here two recent programs, Aggrescan (Conchillo-Sole *et al.* 2007) and FoldAmyloid (Garbuzynskiy *et al.* 2010) are described.

The Aggrescan program (Conchillo-Sole *et al.* 2007) is based on the assumption that short (5-11) residue regions in a protein sequence called “hot-spots” can nucleate fibril formation. Consequently, if a protein sequence contains a hotspot it is considered amyloidogenic. Aggrescan was developed with experimental data from an in vivo system using the 42 amino acid human peptide amyloid-β (Aβ-42) (de Groot *et al.* 2006). This system attaches a green fluorescent protein (GFP) 12 residues upstream to the Aβ-42 region. It was shown that in some cases *Escherichia coli* cells express high levels of this fusion protein but show very little fluorescence. It is believed that this is because the formation of fibrils interferes with the correct folding of GFP and hence reduces the emission of fluorescence.

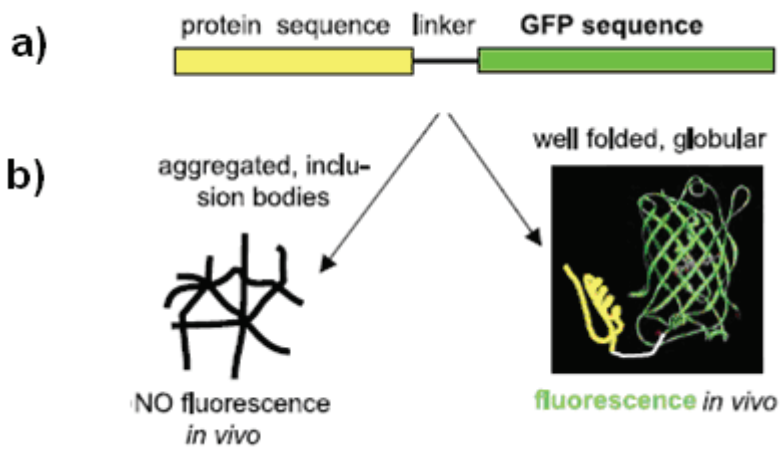


Figure 3. *The in-vivo system. Figure adopted from (Wurth et al. 2002).*
A. The Aβ-42 GFP fusion protein is expressed in Escherichia coli.
B. High aggregation of the Aβ-42 region leads to low fluorescence as it competes with the formation of correctly folded GFP structure. Inversely, low aggregation gives high fluorescence.

The Aβ-42 peptide contains a central hydrophobic region Leu17-Val18-PHE19-PHE20-ALA21 which is considered important to aggregation (de Groot *et al.* 2006). Residue 19, in particular, has been shown to affect fibril formation. Position 19 was mutated to all 19 other amino acids and the *in vivo* system was used to determine their effects on amyloidogenicity. This created the aggregation propensity scale. When a sequence is entered into the program each residue has an amino acid aggregation propensity value assigned to it. A sliding window of 5, 7, 9, or 11 residues is then passed through the sequence, the average aggregation propensity value (**aapv**) is calculated, and then assigned to the central residue. The hot spot threshold (HST) is a predetermined value that corresponds to the average of the 20 naturally occurring amino acids weighted by their frequencies in the Swiss-Prot database. A "hot spot" is a region of the sequence that contains five or more consecutive residues that have an **aapv** higher than the HST and does not have proline residue.

FoldAmyloid (Garbuzynskiy *et al.* 2010) also uses the assumption that short stretches of 5 residues each are vital to the amyloidogenic potential of a sequence. In this case the aggregation propensity scale is determined by the statistical analysis of the known 3D structures of globular proteins. It was shown that two characteristics co-relate well with amyloidogenicity: expected probability of hydrogen bond formation and expected

packing density. FoldAmyloid also assesses the backbone hydrogen bond propensity in terms of acceptors and donors. It can be used with each scale separately or a hybrid scale that combines all three. The program uses a sliding window method similar to Aggrescan to determine the amyloid forming regions of a given sequence.

To develop the program a database of 3769 proteins was constructed. To ensure that the database was representative of all kinds of structures it was constructed to contain structures that were all-α structure, all-ß structure, or a combination of both. To calculate packing density, a residue was considered to be in contact if its non-hydrogen atoms were within 8Å of another residue. Neighbouring residues were excluded from this analysis. The packing density of each amino acid was calculated as the ratio of contacts observed for that amino acid over the total number of times it occurs in the database. To calculate hydrogen (H) bonds four variants were considered: backbone-backbone, backbone-sidechain, sidechain-backbone, and sidechain-sidechain. Backbone-backbone H-bonds were found using the DSSP program. The others were found using a program developed by the authors which uses geometric criteria (distance and angle of hydrogen bond). H-bonding potentials were then calculated for each of the 20 amino acids by dividing the number of times an amino acid was found to be taking part in a hydrogen bond by the number of times it occurs in the database.

1.2.2 Evaluation of properties of β-structural conformation

The major building blocks of amyloids are β-strands, which have an extended conformation with conserved apolar and variable (generally polar) residues alternating along the chain. A number of methods use this information to improve the prediction of amyloidogenic regions.

One of them is the Zyggregator method that takes into consideration patterns of 7 or more residues with alternating apolar and polar residues (Tartaglia and Vendruscolo 2008). To calculate the aggregation propensity, this method also uses a set of physico-chemical properties of amino acid residues such as hydrophobicity, charge, and the propensity to adopt α-helical or β-structural conformations. These properties were derived by fitting the expression used to calculate the aggregation propensity on a database of mutational variants for which aggregation was measured *in vitro* (Chiti *et*

al. 2003; DuBay *et al.* 2004). Zyggregator also considers the flanking residues ("gatekeeper" residues) of a given sliding window for the presence of charged residues of the same sign, as this may reduce aggregation by electrostatic repulsion. In a majority of cases a polypeptide chain should be unfolded to aggregate. Therefore, when applied to structured proteins, prediction methods need to estimate probability of the protein or parts of it to be unstructured. Zyggregator has this option, evaluating the local stability of protein structure by CamP program (Tartaglia *et al.* 2007).

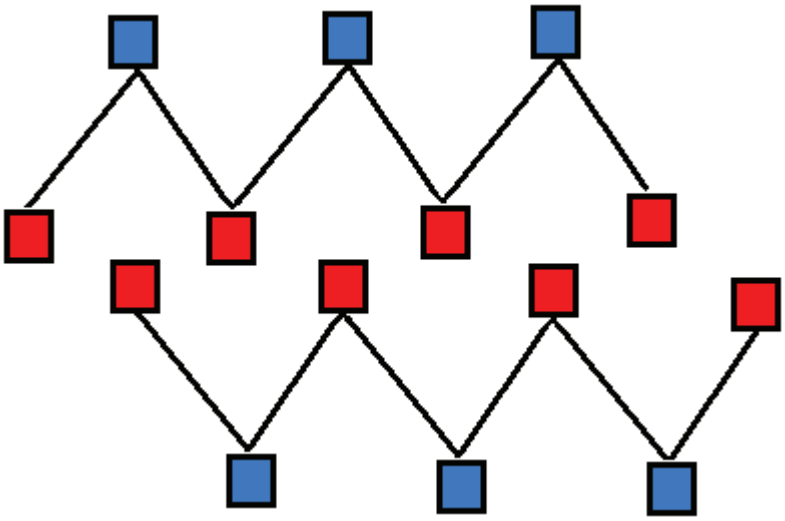


Figure 4. *The alternating pattern of polar and apolar residues taken into consideration by Zyggregator shown in blue and red respectively.*

The TANGO predictor of β-structural aggregation (Fernandez-Escamilla *et al.* 2004) uses a statistical mechanics approach to make secondary structure predictions. For a given sequence this method considers different competing conformations (random coil, β-turn, α-helix, and β-sheets) and predicts which is most likely to occur. The algorithm is based on the following assumptions: (i) a particular amino acid sequence is aggregation-prone if it has high propensity to form β-structure, (ii) all residues of the β-region are buried in the hydrophobic interior of the aggregate, (iii) complementary charges in the selected window establish favourable electrostatic interactions, and (iv) the overall net charge of the peptide disfavours aggregation. TANGO considers that peptides have a tendency for aggregation when they possess segments of at least five consecutive residues in the predicted β-aggregate conformation. Zyggregator and

TANGO both take into account the effect of physico-chemical conditions such as pH, temperature, ionic strength, and the trifluorethanol concentration on aggregation.

1.2.3 Assessment of the pairwise side-chain to side-chain interactions within β-sheets

A β-strand can not exist on its own. It is stabilized only by interaction with other β-strands. The main source of stabilization is by the formation of hydrogen bonds along the main chain. However, side-chain to side-chain interactions between them provide sequence specific stability as well. A variety of ways have been developed to determine the propensity of interaction between side-chains within β-sheets. Two representative examples are the PASTA program (Trovato *et al.* 2007) and the BETASCAN program (Bryan *et al.* 2009).

The central component of the PASTA predictive algorithm (Trovato *et al.* 2007) is the energy calculations for pairs of amino acids interacting via their backbones in β-strands. A non-redundant set of globular proteins was analysed to count the pairs of amino acids that form contacts (C-α atoms lie within 6.5Å of each other) between the β-strands of β-sheets. This analysis was conducted separately for parallel and anti-parallel β-strands. This contact occurrence data is then used to calculate pairwise scores using a Boltzmann distribution. The scores are used to predict the localization and the preferred 3D conformation (parallel or anti-parallel, shifted, or in register) of a given protein.

The BETASCAN also relies on beta pairing propensities but it focuses primarily on the parallel orientation of β-strands since they occur the most frequently. The program determines the potential of parallel β-strands to be formed based on the observed preferences of each pair of residues in parallel β-strands to be hydrogen bonded. To determine these preferences a database of non-redundant structures were taken from the Protein Data Bank. Next the STRIDE algorithm (Frishman and Argos 1995) was used to find β-sheets with solubility differences between its two faces (amphipathic β-sheets). It uses torsion angle and hydrogen bond strength analysis of proteins to determine the secondary structures they can form. The likelihood of a sequence to form parallel β-strands is determined by its propensity to form β-strands multiplied by its propensity to form β-strands. It uses a hill-climbing algorithm to determine if rotation of the β-strands

by 180°, addition or subtraction of residues to the fibril forming region, or shifting the first or second β-strand pairs can give rise to more likely to form β-strands and hence predicted to be more amyloidogenic.

1.2.4 Methods inspired by the understanding of the amyloid structures of short peptides

Since 2005 several crystal structures have revealed for the first time the side chain interactions between β-sheets of short peptides (Nelson *et al.* 2005; Sawaya *et al.* 2007). The micro crystals analysed have the following sequences: GNNQQNY and NNQQNY. They are from the sup35 protein of *Saccharomyces cerevisiae* and form the "cross-β spine." The basic template for it is two parallel β-sheets oriented anti-parallel to each other with an interface created by the like-sides of each sheet Figure 5.

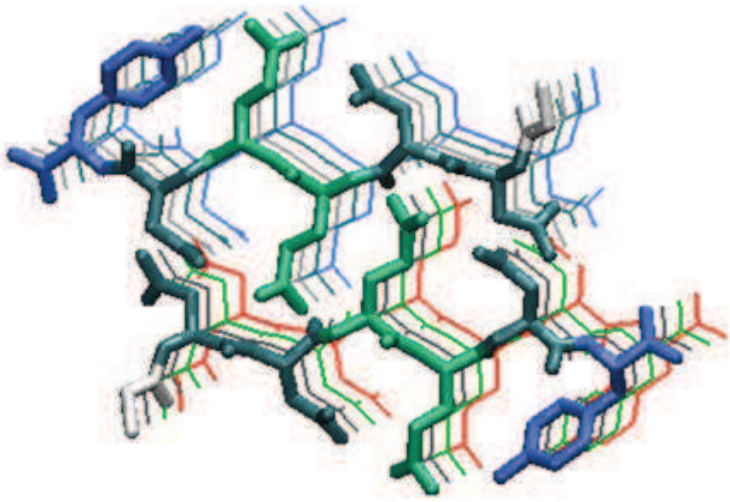


Figure 5. *Interactions of the GNNQQNY fragments within the crystal structure of amyloid-like micro crystals (Nelson, Sawaya et al. 2005)*

The 3D profile method (also known as Zipper DB) (Thompson *et al.* 2006) uses the NNQQNY as a profile or template to determine the amyloidogenicity of sequence data. Initially, a database of six residue peptides called AmylHex was compiled from the literature. It contains 158 peptides, 67 of which are amyloidogenic. 2511 near native templates were made using the sequences in AmylHex and the structure of the

NNQQNY peptide. The program analyzes 6 residue fragments by mapping them onto these templates to create a "profile," which is energetically evaluated using the ROSETTADESIGN program (Simons *et al.* 1999; Liu and Kuhlman 2006). The fragment is considered amyloidogenic if the energy assigned to it is below a predefined threshold.

A similar program, called the Statistical Potential Method here (Zhang *et al.* 2007), also uses the 3D templates generated by small displacements of the crystal structure of the NNQQNY peptide (Nelson *et al.* 2005). However, residue based statistical potential calculations rather than ROSETTADESIGN analysis is used to evaluate the energy of the sequences mapped onto these templates.

Another program called Waltz (Maurer-Stroh *et al.* 2010) uses an expanded version of the AmylHex dataset (Thompson *et al.* 2006) as a learning set to determine a position specific scoring matrix (PSSM) to identify amyloid forming sequences. The PSSM analysis is augmented with a physical property term that combines 19 physical properties of amino acids known to correlate with amyloid formation, and a position specific pseudo energy matrix derived from the mutational analysis of the sup35 GNNQQNY peptide (Nelson *et al.* 2005). The PSSM was motivated by realization that the analysis of amino acid composition alone does not take into account all the information that was available at the time of its development. The position of a given amino acid within the fibril is also an important factor. So the PSSM was made to determine whether certain residues had specific preferences for different positions in the six residue motif (figure 6). Each cell in matrix represents the beneficial or detrimental effect a given natural amino acid at a given position has to fibril formation, (For example, the effect of Ala at position 1, or the effect of Leu at position 5).

.

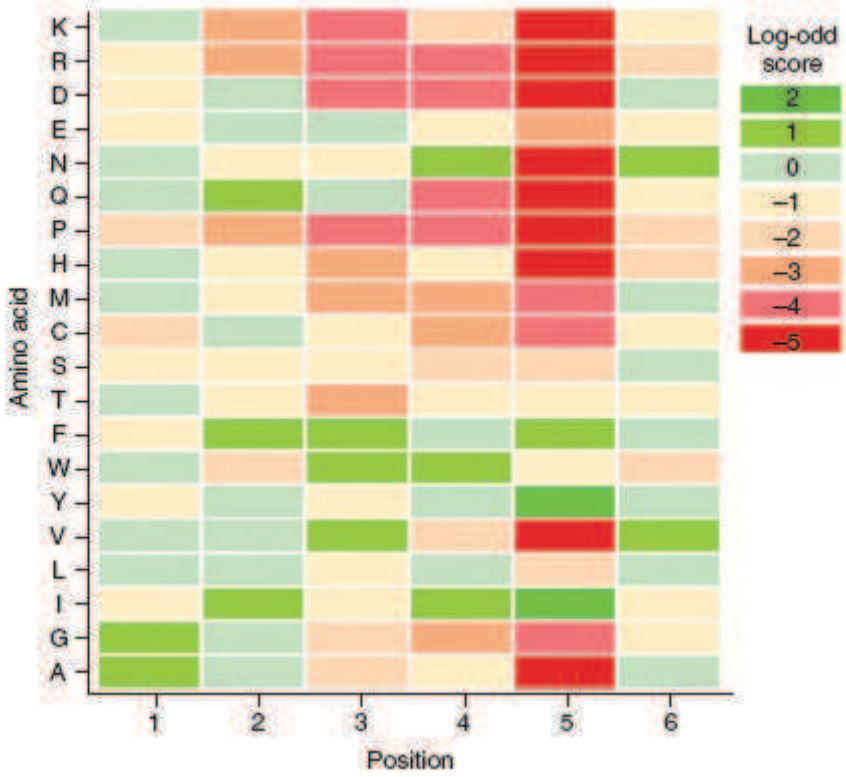


Figure 6. *Position specific scoring matrix for natural amino acids determined using the AmylHex database. If an amino acid at a specific position is favourable for fibril formation it is shown in green. Otherwise it is red. The scale on the right shows the colours for intermediate values. Adopted from (Maurer-Stroh et al. 2010).*

Waltz also uses the physical property descriptors for B sheet forming propensity, A helix forming propensity, and solvation to enhance its predictive abilities. A list of roughly 700 parameter sets was whittled down to 19 properties with the highest predictive strength (Maurer-Stroh *et al.* 2010).

Finally, the crystal structure of the GNNQQNY sup35 fragment (Nelson *et al.* 2005) was reduced to poly-alanine and then mutated to all possible combination of naturally amino acids. Energy estimations using the FoldX (Guerois *et al.* 2002) program were then used to make the position specific pseudo energy matrix.

1.2.5 Estimations of the probability of structured proteins to become partially unfolded

To form cross-β amyloids, a polypeptide chain with high amyloidogenic potential needs to be unstable within its native 3D structure or be completely unfolded. Indeed, experimental studies show that most of the known amyloid-forming sequences (for example, amyloid-β, α-synuclein, Ure2p, and Sup35p) are unstructured in their non-amyloid state. Proteins that fold into soluble 3D structures may also contain a number of amyloidogenic regions hidden in their structures. Significant efforts have been dedicated to the identification of such hidden regions (also known as 'conformational switches' or "chameleon" sequences) within globular proteins that are innocuous in their normal state (Chiti *et al.* 2000).

Some methods developed for prediction of amyloidogenicity address this problem. For example, the Zyggregator method includes an option to evaluate the local stability of protein structure (Tartaglia *et al.* 2007). The Net-CSSP method (contact-dependent secondary structural propensity) (Yoon and Welsh 2004; Kim *et al.* 2009) quantifies the influence of tertiary interactions on secondary structure preference by using an artificial neural network-based algorithm and seeks to find short regions with a hidden potential to form β-sheets.

Another web-based tool, Amylpred, combines the results of amyloidogenicy predictions with the SecStr secondary structure prediction tool (Hamodrakas *et al.* 2007). The SecStr tool uses five different methods of the secondary structure prediction. If, according to the secondary structure prediction, the amino acid stretches have ambivalent propensities for α-helix and β-strand, they are considered as regions with the potential 'conformational switches'. After that several approaches such as, FoldAmyloid (Garbuzynskiy *et al.* 2010) and scanning of proteins with amyloidogenic motif extracted from the known fibril-forming peptides (Lopez de la Paz and Serrano 2004) are applied to the sequence. Regions of the structured protein that are simultaneously identified as the 'conformational switches' and highly amyloidogenic considered to be the amyloidogenic determinants.

1.3 Evaluation of Prediction Methods

To evaluate prediction methods, benchmark datasets of amyloid-forming and non-forming sequences are required. When doing so, the primary problem is the limited number of known amyloid-forming proteins. Today, only about 20 amyloid-forming proteins are known to be linked to diseases (Pepys 2006). Although it is true that the datasets can be enriched by adding known mutants of these proteins, this does not solve the problem, as the datasets become biased towards certain overrepresented sequences. Moreover, whereas prediction methods are designed to exclusively detect cross-β amyloids, disease-related fibrils are heterogeneous in terms of their 3D structure. Some are formed by stacks of native or refolded globular structures, (Westermark *et al.* 1990; Elam *et al.* 2003; Sanders *et al.* 2004) and do not necessarily exhibit cross-β structure. Care must also be taken when developing the negative set. It is tempting to use globular proteins as they are soluble and non-amyloidogenic. Most prediction programs, however, operate using only sequence information, and will incorrectly predict amyloidogenic candidates that are in fact hidden inside the protein structure. Furthermore, when one considers that different amyloid-forming proteins form fibrils at different conditions (concentration, ionic strength, pH, etc) it becomes evident that the task to construct testing datasets of high quality is extremely challenging.

Most of the methods use datasets of short peptides. The reasons are that short peptides can be synthesized easily and tested in the same or similar experimental conditions for the formation of amyloid fibrils. Moreover, soluble short peptides can be used directly as a non-amyloidogenic set. As these peptides are unfolded, they do not have the problem of structurally hidden regions found in folded proteins. Finally, the usage of short peptides is in agreement with the predominant paradigm underlying existing prediction algorithms: short (about 6 residue) regions are sufficient for forming amyloid fibrils of full-length proteins.

There are several popular benchmark datasets of short peptides. The first large dataset was compiled for the testing of the TANGO algorithm (Fernandez-Escamilla *et al.* 2004) and consisted of 78 amyloidogenic and 172 non-amyloidogenic peptides mostly from human disease related proteins. Peptides were considered to be aggregating when their circular dichroism or NMR spectra had concentration dependence in the range

between 1 mM and 5 mM, or when binding to an amyloid-reporting dye (ThT) was observed. Another set of experimentally determined amyloid-forming peptides was selected from the literature and used to test AGGRESCAN program (Conchillo-Sole *et al.* 2007). The most frequently used data set is AmylHex. It contains 158 six-residue peptides of which 67 have been shown to form fibrils and 91 are soluble (Thompson *et al.* 2006). A majority of the dataset consists of mutants of STVIIE peptide, as well as hexapeptides and their mutants from amylin, tau, insulin, β2-microglobulin. Recently, the AmylHex dataset was supplemented by 49 new amyloid-forming and 71 non-amyloid-forming hexapeptide sequences (Maurer-Stroh *et al.* 2010) to bring the total number of amyloid forming hexapeptides to 116 positive and 103 negative sequences. Several other predictors of amyloidogenicity used one of the datasets mentioned above or their combinations.

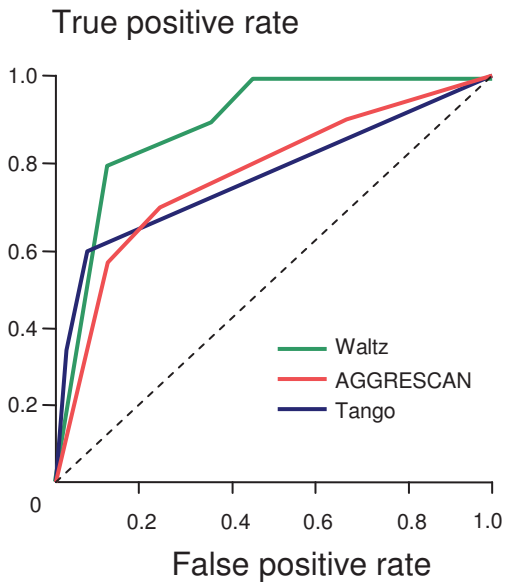


Figure 7. *Benchmarking of TANGO, AGGRESCAN, and Waltz on the combined dataset.*

Figure 7 shows our benchmarking results for three programs (TANGO, AGGRESCAN and Waltz) on a combined set of the sequences from all the datasets mentioned above. The tested programs display good results, correctly identifying 65%, 71% and 80% of the amyloid-forming peptides, correspondingly, and having only 17%, 25% and 15% of false positives in the set of non-amyloidogenic peptides. Waltz performs better than the other programs, however, it is necessary to remember that a large number of peptides from the combined dataset were used by this program as a training set (Maurer-Stroh *et al.* 2010).

The other approach typically used to demonstrate the power of the methods was the prediction of known pathogenic or protective mutants of amyloid-forming proteins to demonstrate the ability to predict the observed change in the amyloidogenicity (Fernandez-Escamilla *et al.* 2004; Conchillo-Sole *et al.* 2007). In addition, the programs are tested for the prediction of locations of amyloid-forming regions in longer peptides (30-40 residues) and full-length proteins. Especially those, with a natively unfolded monomeric state, and experimentally verified locations of amyloid forming regions (Figure 8). The most frequently used examples for such tests are amyloid-β, α-synuclein and amylin. In Figure 8, the predictions of amyloidogenic "hot spots" in fibril-forming regions of amyloid-β and Het-s prion are shown. The programs generate satisfactory predictions for amyloid-β peptide, while in the Het-s prion region, the predictions are less credible. For example, Waltz program does not find any amyloid-forming region within the Het-s prion domain. This can be explained by the absence of the Het-s peptides in its training set, or by some differences of the Het-s fibril structure from the typical cross-β amyloids. The amyloid-β structure represents a stack of identical peptides, but the Het-s cross-β fibril is formed by the repetitive element with two slightly different β-strands alternating along the fibril axis.

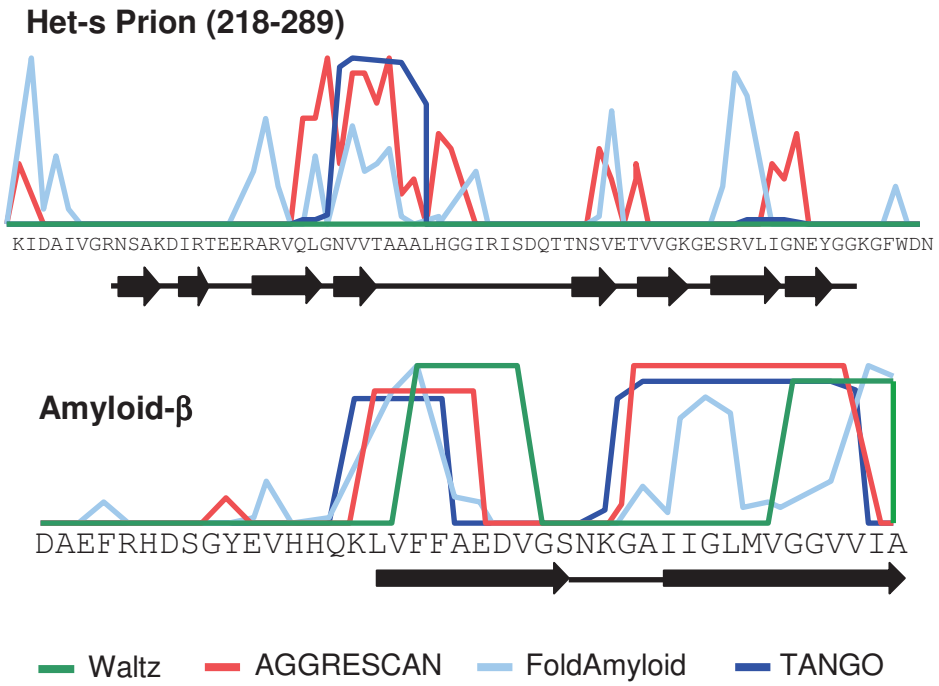


Figure 8. *Results of Waltz, AGGRESCAN, FoldAmyloid and TANGO when tested for the prediction of locations of amyloid-forming regions in longer peptides (30-40 residues) and full-length proteins.*

1.4 Understanding the 3D Structure of the Amyloid

Structural information on amyloids comes from three major sources. None of them provides complete structural data; however, these insights can be combined to produce models for fibrils. The sources are:

- Experimental techniques that give incomplete information about atomic structure of amyloids
- X-ray crystallography of short peptide fragments in amyloidogenic states,
- X-ray crystallography of the β-solenoid structures.

Initial details were determined by traditional experimental techniques. X-ray diffraction provided some of the earliest clues about the overall structure of fibrils. It established the cross-β pattern of fibrils (Astbury *et al.* 1935; Eanes and Glenner 1968). Electron Microscopy (EM) and Atomic Force Microscopy (AFM) provided nanometer resolution of the ultrastructural characteristics of amyloids such as fiber length, width, and morphology (curvature, periodic twists and surface characteristics). EM was used to determine the long, unbranched, "straight" nature of the fibrils, the typical fiber width of 5-15 nm, the periodic twist, and to conclude that many amyloid fibrils are made of the bundling together of thinner protofibrils (Cohen and Calkins 1959; Boere *et al.* 1965; Shirahama and Cohen 1965). Scanning transmission electron microscopy has been used to determine mass-per-unit length of amyloids (Sen *et al.* 2007). Cryo-EM has been used to make several different models (Jimenez *et al.* 1999; Jimenez *et al.* 2002; Meinhardt *et al.* 2009). Tilted-beam transmission electron microscopy, EM, and AFM have been used to shed light on how intermediates in the aggregation pathway progress to fibrils (Goldsbury *et al.* 2000; Chen *et al.* 2009). Spectral techniques such as Fourier Transform Infrared Spectroscopy and Circular Dichroism can provide an estimation of the contribution of β-sheets, α-helices, or loops to the structure. They have been used to confirm the high β-sheet content of the fibrils and to determine the different concentrations of β-structure in different fiber preparations of the same protein (Termine *et al.* 1972; Gasset *et al.* 1993). Proline mutations have been used to determine regions of β-structure since they are β-sheet breakers (Williams *et al.* 2004). Mutations to cysteine can be labelled with a paramagnetic spin label for Electron Paramagnetic Resonance. This indicates the presence or absence of structure in this

region and can be used to measure the intra and intermolecular distances between probes (Serag *et al.* 2002; Torok *et al.* 2002; Chen *et al.* 2007). However, a drawback of this method is that the mutations may change the part of the structure being examined (Toyama and Weissman 2011). Finally, solid state NMR (ssNMR) has been used to differentiate between parallel (in register) and anti-parallel structures, and to resolve the locations of the β-strand regions and the unstructured loops. (Jaroniec *et al.* 2004; Iwata *et al.* 2006; Shewmaker *et al.* 2006; Luca *et al.* 2007; Shewmaker *et al.* 2009). It can also be used to find the details of the structure in highly ordered fibrils as in the HET-s protein (Siemer *et al.* 2005; Van Melckebeke *et al.* 2010).

In 2005, the structure of micro crystals formed by the sup35 protein of *Saccharomyces cerevisiae* was realised (Nelson *et al.* 2005; Sawaya *et al.* 2007). For the first time the interactions of side-chains within the core of the fibril were revealed. They were very tightly packed into a "cross-β spine", an arrangement with extensive interdigitation between side chains. Since then several other structures of short peptides (~6 residues) engaging in amyloid-like fibrils have been resolved (Sawaya *et al.* 2007). The discovery of the cross-β spine showed that short peptide could provide important information.

Finally, the structure of amyloids was further elucidated by studies on a class of proteins called β-solenoids which are based on solenoidal winding of β-structural units (Kajava and Steven, 2006). A large number of solenoid 3D structures have been resolved and the detailed analysis of their standard conformations conducted. These structures are the closest known template for amyloids. This helped reveal the conformations adopted in the loop regions of solenoids and by doing so helped understand the structure of amyloid fibrils linked to major human diseases (Hennetin *et al.* 2006).

Based on the experimental information, several models for amyloid fibrils were constructed (Thakur and Wetzel 2002; Der-Sarkissian *et al.* 2003; Govaerts *et al.* 2004; Kajava *et al.* 2004; Margittai and Langen 2004; Kajava *et al.* 2005; Krishnan and Lindquist 2005; Luhrs *et al.* 2005; Ritter *et al.* 2005; Sikorski and Atkins 2005; Baxa *et al.* 2006; Ferguson *et al.* 2006; Inouye and Kirschner 2006; Nelson and Eisenberg 2006; Petkova *et al.* 2006; Shewmaker *et al.* 2006; Luca *et al.* 2007; Andronesi *et al.* 2008; Jeganathan *et al.* 2008; Wasmer *et al.* 2008; Wiltzius *et al.* 2008). Recently, it was shown that a majority of structural models of naturally occurring and disease-related

amyloid fibrils can be reduced to a so called "β-arcade" (Kajava *et al.* 2010). Each β–arcade has a double-layer structure in which 2 parallel in-register β–sheets face each other creating a columnar structure. The side chains protrude into the space between apposing β-sheets to form tight inter-digitated packing. They are produced by stacking of β-strand-loop-β-strand motifs called "β-arches" (Figure 9).

A majority of globular structures contain strand-loop-strand motifs called β-hairpins. In these structures the strands form an anti-parallel β-sheet (Figure 9). In the β-arch each strand is relatively rotated ~90° so that they interact via their side chains (Baxa *et al.* 2006).

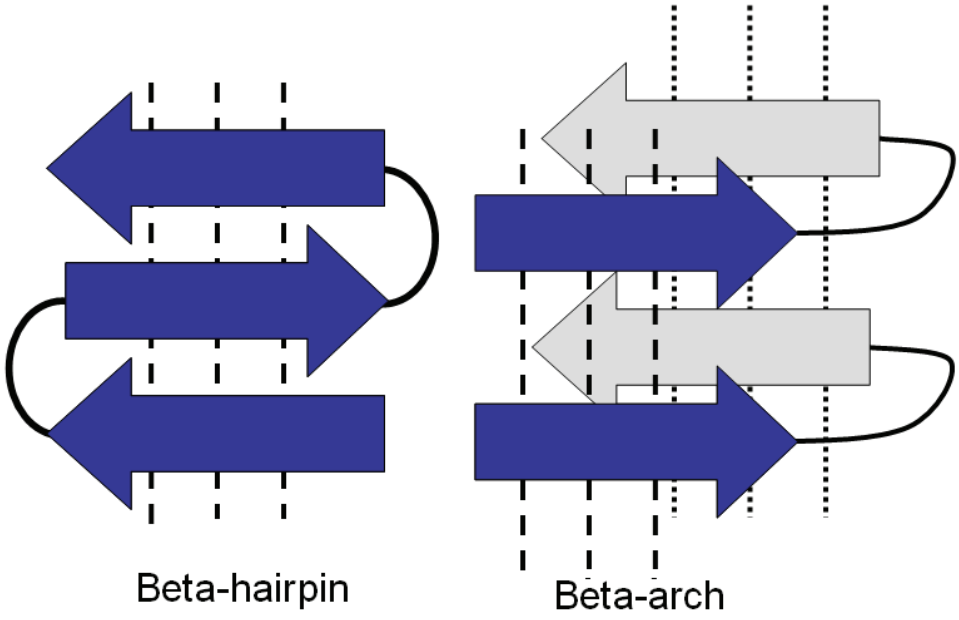


Figure 9. *A comparison between hairpins and arches. The arrows represent β-strands that interact via H-bonding (shown by dotted lines).*

Amyloid fibrils consist of one or several protofibrils built of β-arcades (Figure 10). Topologically, they are of three types of models for fibrils (Kajava *et al.* 2010). The first type is typified by protofibrils of Amyloid-β, the K3 fragment of B2-microglobulin, human amylin, and CA150 protein (Luhrs *et al.* 2005; Ferguson *et al.* 2006; Iwata *et al.* 2006; Petkova *et al.* 2006; Luca *et al.* 2007). They are composed of structural units composed of one β-arch that are stacked on top of each other along the fibril axis and form a double layer of parallel β-sheets. The second type corresponds to protofibrils

proposed for the structures formed by Ure2p, Sup35, α-synuclein, poly-Gln tracts, amylin tau, and the B1 domain of the IgG binding protein G (Der-Sarkissian *et al.* 2003; Kajava *et al.* 2004; Margittai and Langen 2004; Kajava *et al.* 2005; Wang *et al.* 2005; Wiltzius *et al.* 2008). In this case each polypeptide chain has several β-sheets that are connected by loop and it zigzags to create a planar serpentine fold. These serpentine are stacked upon each other axially in register, thus forming an array of parallel β-sheets within a so called super-pleated β-structure. The third type of protofibils applies to the HETs-prion. The HETs-prion is composed of two coil β-solenoids stacked on top of each other (Wasmer *et al.* 2008).

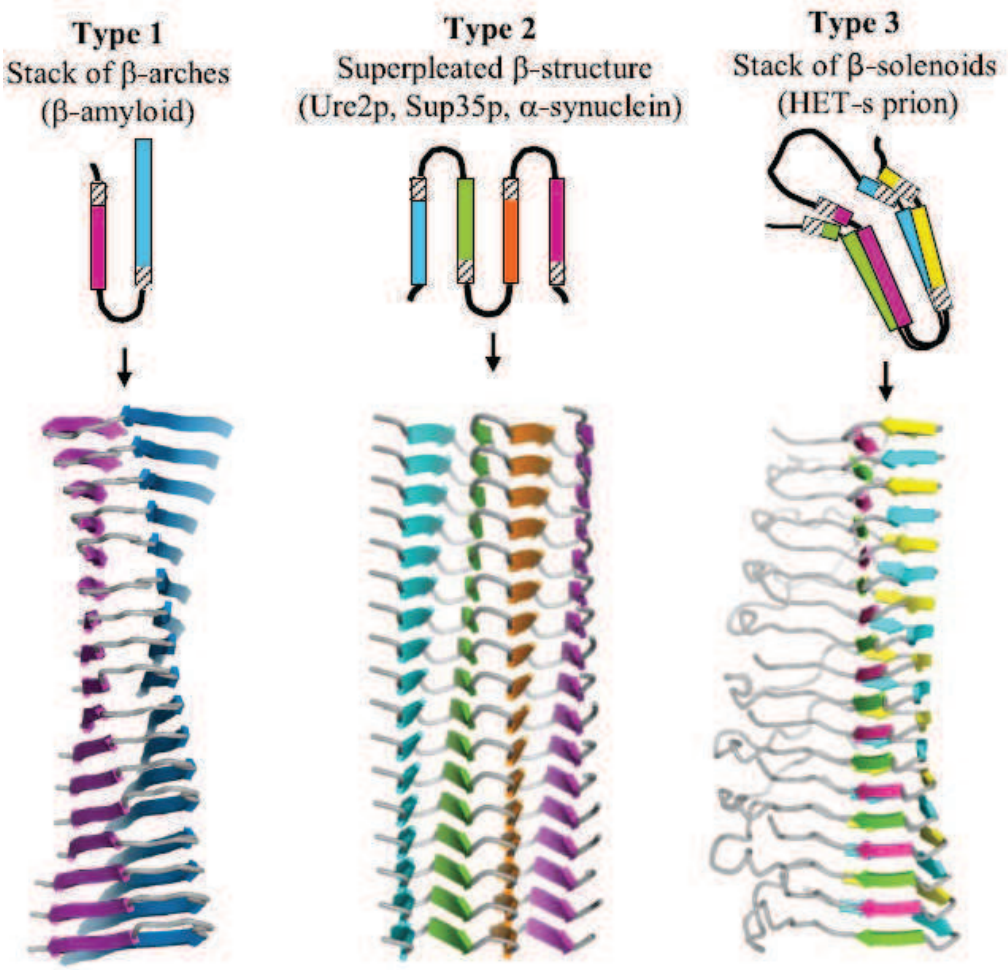


Figure 10. *Three types of models for amyloid fibrils. Type one is composed of identical β-arches stacked on top of each other. Type two is made by the stacking of planar serpentine folds. In type three the repeating unit is two coiled β-solenoids. Adopted from (Kajava et al. 2010)*

2. Formation of Objectives

Several computational methods have been developed to predict the propensity of polypeptides to form amyloids based on sequence analysis. Many of the methods have rendered excellent performance capabilities in the numerous tests. These algorithms use the assumption that a short sequence (about 6 residues) is sufficient to trigger the amyloid formation of a given protein. Consequently, they achieve their best results among short peptides. However, the analysis of short peptides is largely un-equivalent to the *in vivo* formation of disease related amyloids. Indeed, peptides of less than about 15 residues rarely reach fibril-forming concentrations in human cells, as once produced, they are rapidly degraded by endogenous proteases (Saveanu *et al.* 2002). Although it is true that a short fibril-forming region may occur within a longer polypeptide chain, fusion of short amyloidogenic peptides with soluble proteins has not yielded convincing results, only triggering fibrillation at very high concentrations (Esteras-Chopo *et al.* 2005; Guo and Eisenberg 2008). Additionally, known naturally occurring amyloid-forming proteins have amyloidogenic regions that are longer than 15 residues. Finally, recent experimental techniques reveal that the minimal structural element of the majority of disease-related amyloid fibrils is a columnar structure produced by stacking of β-strand-loop-β-strand motifs spanning over 15-20 residues.

Current programs for amyloid prediction are unable to make use of the full ensemble of recently obtained structural information. The objective of this work was to fill this void and to develop a new approach based on the assumption that sequences that are able to form β-arcades are amyloidogenic. Next, in the Results section the development of the algorithm and a computer program called ArchCandy is explained.

3. RESULTS

3.1 Building a dataset for Naturally Occurring Amyloids and Benchmarking of Existing Programs for Amyloid Prediction

Most amyloid prediction programs use the paradigm that short, 6 residue long peptides are sufficient to initiate fibril formation. The datasets used to test them are derived from the *in vitro* analysis of hexapeptides and it was demonstrated that these programs accurately predict short amyloid-forming peptides (Fernandez-Escamilla *et al.* 2004; Conchillo-Sole *et al.* 2007; Maurer-Stroh *et al.* 2010; Ahmed and Kajava 2013). However, it must be emphasised that the eventual goal for all methods is the correct prediction of amyloid fibril formation in naturally occurring and disease-related proteins and peptides. Amyloid forming sequences involved in diseases (Pepys 2006) tend to be longer in length. To test the performance of existing programs on these naturally occurring sequences a new dataset was derived from literature (Ahmed and Kajava 2013). It is composed of proteins or peptides known to form amyloids *in vivo* that were taken from scientific publications with the following criteria: their amyloidogenic regions are unfolded in their native state, and they form cross-β fibrils *in vivo* or under conditions that are close to the physiological (pH 5.5-7.5, concentration of protein up to 150 μM). This dataset contains 23 sequences from a diverse array of sources (Table 3). Human proteins and peptides are represented by sequences related to disease (e.g. Amyloid-β, α-synuclein) as well as functional proteins (PMEL17). Bacterial or fungi proteins are represented by functional amyloids (e.g. Chaplin proteins from *Streptomyces coelicolor*, Curli proteins in *Escherichia coli* and Prion Formation Protein 1 from *Saccharomyces cerevisiae*). The negative set was extracted from the DisProt database of disordered proteins (Vucetic *et al.* 2005) with the following criteria: sequences are disordered in their entirety and have less than 150 residues. The negative set contains 52 sequences (Annex II).

Table 3. The positive set of 23 naturally occurring proteins and peptides known to form amyloids *in vivo.*

Protein or peptide name	Amyloid region length (aa)	Amyloid forming type	References
Human amyloid-β 42	42	Human disease-linked	(Kirschner *et al.* 1986)
Human α-synuclein	140	Human disease-linked	(Giasson 2000)
Human β 2-microglobulin mutant fragment	22	Human disease-linked	(Iwata *et al.* 2006)
Human CA150	40	Human disease-linked	(Becker *et al.* 2008)
Human amylin	37	Human disease-linked	(Fox *et al.* 2010)
HET-s Prion from *Podospora anserina* (218-289)	71	Functional	(Dos Reis 2001)
Human calcitonin	32	Human disease-linked	(Kamihira *et al.* 2000)
Human Semen-derived Enhancer of Viral Infection (SEVI) Fibril Forming peptide of Prostatic Acid Phosphatase Peptide (248-286)	39	Human disease-linked	(Ye *et al.* 2009)
Sup35 from *Saccharomyces cerevisiae* (1-114)	114	Functional	(Baxa *et al.* 2006)
Ure2P from *Saccharomyces cerevisiae* (1-94)	94	Functional	(Baxa *et al.* 2006)
Rnq1p from *Saccharomyces cerevisiae* (153-405)	253	Functional	(Baxa *et al.* 2006)

Human Ataxin Diseases (including Huntingtin disease)	≥20	Human disease-linked	(Perutz *et al.* 2002)
Chaplin F from *Streptomyces coelicolor*	52	Functional	(Sawyer *et al.* 2011)
Microcin E492 from *Klebsiella pneumoniae* (16-99)	84	Functional	(Arranz *et al.* 2012)
Prion Formation Protein 1 from *Saccharomyces cerevisiae* (1-100)	100	Functional	(Santoso *et al.* 2000)
Human RIP1 (519-560)	42	Functional	(Li *et al.* 2012)
Human RIP3 (439-479)	41	Functional	(Li *et al.* 2012)
Human TDP (TAR DNA-binding Protein; 281-332)	52	Human disease-linked	(Chen *et al.* 2010)
Human Prp (23-230)	208	Human disease-linked	(Cobb *et al.* 2007)
murine serum amyloid A-2 protein isoform SAA2.2 (20-122)	103	Disease-linked	(Ye *et al.* 2011)
CsgA from *E. coli* K12 (21-151)	131	Functional	(Shewmaker *et al.* 2009)
CsgB from *E. coli* K12 (22-151)	130	Functional	(Shewmaker *et al.* 2009)
Human Pmel 17 M-α domain (25-467)	443	Functional	(Watt *et al.* 2009)

The performance of existing programs against this dataset is unsatisfactory. They, generally predict a sizable number of false positives (Table 4) when applied to the sequences of longer than 30-40 residues. Other problems of these methods are the over prediction of amyloids in hydrophobic regions, and their poor predictive capability of amyloidogenic sequences rich in polar Gln and (or) Asn. This shortcoming can be explained by the fact that some methods use aggregation propensities values obtained

from the analysis of globular proteins which have the hydrophobic residues as the predominant structure-stabilizing factor.

Table 4. Performance of different methods on datasets of proteins.*

Program**	True positive rate	False positive rate
Waltz	0.666 (12/18)	0.346 (18/52)
Tango	0.277 (5/18)	0.500 (26/52)
Aggrescan	0.722 (13/18)	0.769 (40/52)
FoldAmyloid	0.388 (7/18)	0.750 (39/52)
AmylPred	0.833 (15/18)	0.673 (35/52)

True positive rate: (Number of true positives) / (Total number of amyloid-forming sequences).

False positive rate: (Number of false positives) / (Total number of non-amyloidogenic sequences).

** Tested on a positive set of 18 sequences and a negative set of 52 sequences described in (Ahmed and Kajava 2013).*

*** The default settings of the web-servers were used.*

The performance of existing programs can be summarized thusly. They quite accurately predict short amyloid-forming peptides, and are adept at determining experimentally established fibril-forming regions in full-length proteins. However, they perform poorly on a test set of longer sequences derived from literature. This result revealed imperfection of the previously suggested methods and pointed out the necessity of developing a program that is based on a better performing algorithm.

3.2 Development of an Algorithm for Amyloid Prediction Based on finding β-Arcade Forming Sequences

The unconvincing performance of the existing methods on the one hand and advances in the understanding of the 3D arrangement of the disease-related amyloid fibrils on the other triggered our work on the development of a new method. It has been shown that the core structure of many disease-related amyloids is the β-arcade (Kajava *et al.* 2010). In accordance with this finding, we developed the ArchCandy program to detect protein sequences that are able to form β-arcades. In fact, the name ArchCandy is derived from its function of finding good subsequences or candidates (candies) capable of forming β-arches.

The details of the protein folding into β-arcade structures are largely unknown. Usually, the amyloid fibril formation is preceded by a lag-phase, indicating the presence of a nucleating event and intermediate oligomeric structure(s) (Ma and Nussinov 2002; Marek *et al.* 2010). The importance of the nucleation structure is also confirmed by seeding experiments: where the addition of pieces of an amyloid fibril eliminates or reduces the lag-phase (Harper and Lansbury 1997). Despite the uncertainties in the folding details, the knowledge of the final state – β-arcade structures, provide important information about the probability of the sequences to form amyloids. Therefore, in our algorithm we focused on the evaluation of these final states. For this purpose, first, we needed to get the largest possible set of β-arcade structures (known and modelled) and second, to find a way to evaluate the molecular energy of these β-arcades.

3.2.1 Known and modelled β-arcades

To address the first problem we analysed the known β-arcade structures of amyloids. There are several resolved structures for Amyloid-β (Luhrs *et al.* 2005; Petkova *et al.* 2006; Paravastu *et al.* 2008; Qiang *et al.* 2012) and one each for Human CA150 protein (Ferguson *et al.* 2006) and β2-microglobulin (Iwata *et al.* 2006) (Figure 11).

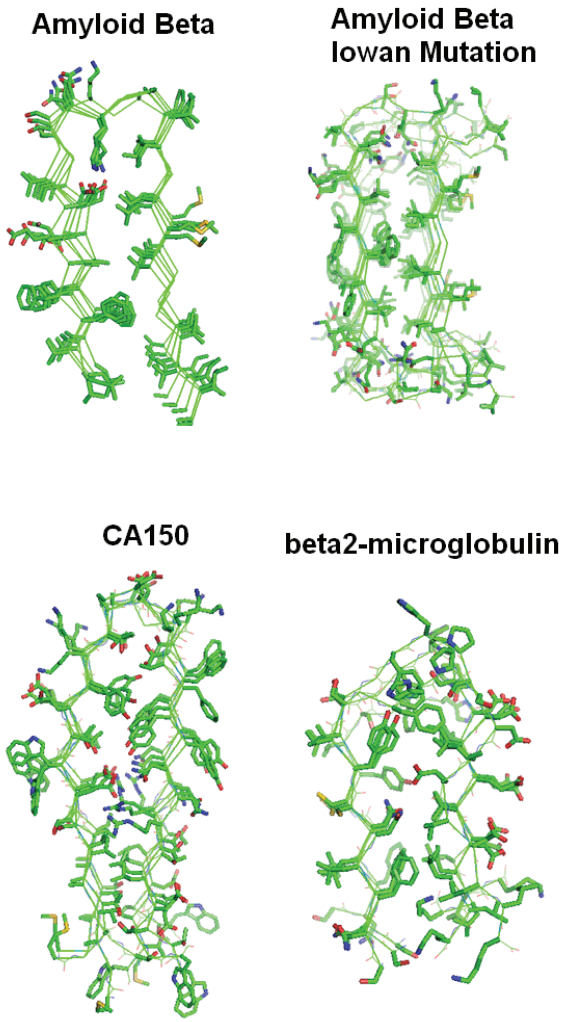


Figure 11. *Some resolved structures for amyloid fibrils. Amyloid-β (PDB code: 2BEG), Amyloid-β Iowan mutation (PDB Code: 2LNQ), Human CA150 protein (PDB code: 2NNT), and β2-microglobulin (PDB code: 2E8D) as visualized by Pymol (Schrodinger 2010).*

The inspection of these structures shows that β-arcades have a well defined boundary between the interior side-chains that form a hydrophobic core and those that are solvent exposed. This boundary is formed by the axial hydrogen bonding between backbones. Polar residues are not suited well to the hydrophobic core and do not occur there often. Gln and Asn are exceptions which are able to form axial hydrogen bonding "ladders" via their side chains. In some cases they are even able to form hydrogen bonds with the backbones of the apposing B-sheet. Charged residues can occur in the hydrophobic region provided they form salt-bridges with oppositely charged residues.

Although the known β-arcade structures provide important insight, they are too few in number to provide sufficient information for the development a program capable of β-arcade discovery. Therefore, to get the more complete set of different β-arcades we used molecular modelling. The main source of polymorphism in β-arches is the length of the β-strands, and the length and conformations of the arc region (Figure 12).

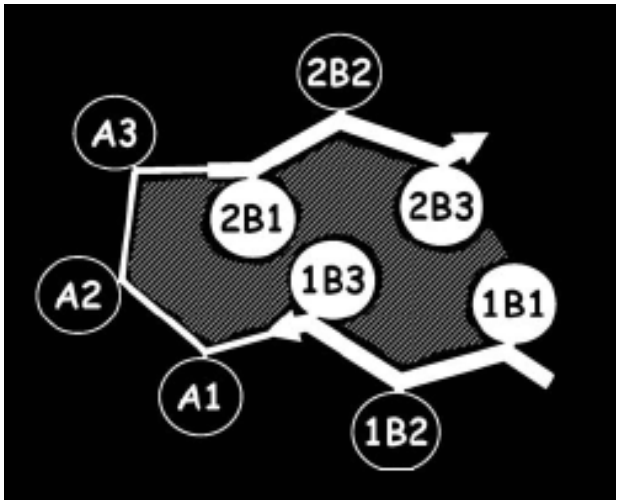


Figure 12. *Nomenclature used to describe residue positions in β-arches. Open and filled circles denote side-chains directed outside and inside the arch respectively. Thick arrows denote β-strands. Shaded region indicates the internal hydrophobic space of the arch. The residues labelled 'A' form the arc region. The number of A-residues varies depending on the arch type. Adopted from (Hennetin et al. 2006)*

It has been shown that U-turns of β-arches flanked by β strand regions are composed of a 3-7 residue long region called an arc (Hennetin *et al.* 2006). The analysis of the known 3D structures of β-solenoid proteins (Kajava and Steven 2006) which contain β-arcs showed that they have a limited number of favourable conformations depending on their length (Hennetin *et al.* 2006). Based on this information, seven template β-arches were used for a set of the modelled β-arcades. Their arcs range between 3-6 residues long. The longest arc region was set to 6 since the analysis of β-solenoids shows that β-arches with arc regions longer than 6 residues are rare in the known 3D structures and are not stacked one over the other but dispersed along the β-solenoids (Hennetin *et al.* 2006). Our modelling also shows that arcs longer than 6 residues start to show high levels of

steric tension when stacked in the long β-arcades. Thus, the representative set of the β-arcs that cover the majority of cases consist of three templates with four residue arcs, two with six residues in this region, and one each with three residues and five residues in the arc (Figure 13).

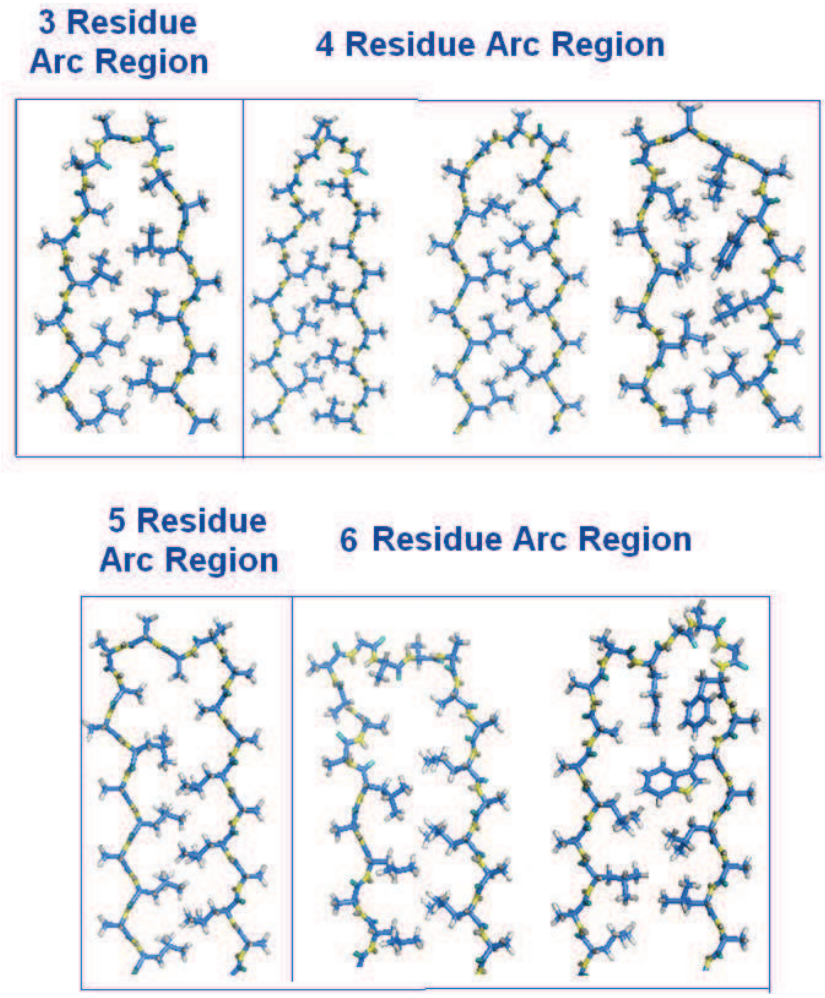


Figure 13. *All seven templates as observed using the Pymol program (Schrodinger 2010)*

The different arc regions with the most frequently adopted conformations were taken directly from known structures of β-solenoids (Hennetin *et al.* 2006). The β-arches were built using the Coot program (Paul Emsley 2010), and were refined by energy minimization using GROMACS (GROningen Machine for Chemical Simulations) (Van Der Spoel *et al.* 2005).

For all β-arch templates the conformation of the β-strand region was identical with Phi= -119 ; Psi=+113 that corresponds to the typical parallel β-structure values (Fraser and MacRae 1973). The β-strands interact through their side chains and the distance between them is ~10 Å. The β-strands were slightly shifted both axially and laterally to ensure maximum inter-digitation and the best knob-to-hole packing of the side chains of the internal residues as was observed in the known structures of amyloid-like crystallites (Nelson *et al.* 2005; Sawaya *et al.* 2007) and β-arcades (Iwata *et al.* 2006; Petkova *et al.* 2006; Paravastu *et al.* 2008; Qiang *et al.* 2012).

It is important to mention that in resolved structures containing stacks of β-arches or β-strands of the same molecule the equivalent internal side-chains have the same rotamers in every β-arch in the fibril (Luhrs *et al.* 2005; Nelson *et al.* 2005; Iwata *et al.* 2006; Kajava and Steven 2006; Petkova *et al.* 2006; Sawaya *et al.* 2007; Paravastu *et al.* 2008; Qiang *et al.* 2012). Despite this constraint the correct prediction of the side-chain rotamers of a β-arch that can occur inside of the β-arcade is still a challenge. To address this problem the optimal rotamers adopted by the internal residues were determined by evaluating possible rotamers manually and by energy evaluations implemented in GROMACS. To create β-arcades from these β-arch templates an in house program called Arch3D was written in Java (http://www.java.com/en/). It has the ability to axially stack a given β-arch in a parallel and in-register manner with user defined axial displacement and twist. Fibrils are known to not be completely flat, but slightly left-hand twisted when viewed along the axial axis. In our analysis all structures have an axial shift of 4.8 Å between β-arches (optimal distance to form axial hydrogen bonds) and a twist of 0.5° that occurs in the known β-structures (Fraser and MacRae 1973). Energy minimization was then applied to the built structures to refine the stereochemistry of the polypeptide chain and remove close contacts. GROMACS program was able to maintain the property – "the equivalent internal side-chains have the same rotamers in every β-arch" during energy minimization.

3.2.2 Choosing an approach to evaluate the probability of β-arcades formation

The next task was to choose a way to evaluate the energy of known and modelled β-arcades. The most obvious way to evaluate the probability of β-arcade formation is using existing programs that calculate molecular energies. To assess the quality of these programs they were tested, on one hand, on sets of known β-arcade fibrils and on the

other, fibrils that from general consideration are unlikely to be formed. For this study several of the most popular programs were tested such as GROMACS version 3.3.4 (Van Der Spoel *et al.* 2005), Rosetta, FoldX, and the energy minimization module of Modeller (Simons *et al.* 1999; Fiser *et al.* 2000; Guerois *et al.* 2002; Eswar *et al.* 2007) . The β-arcades built by Arch3D were subjected to energy minimization by the corresponding program followed by evaluation of the molecular energy of the minimized structure. However, our tests revealed that energy evaluation alone is insufficient to completely assess the structures.

Here are some examples of contradictory results:
The introduction of a charged residue into the β-strand region in a solvent exposed position will have very little effect. However, if the same mutation is made at an inside position in the β-strand it can even completely prevent fibril formation. In this hypothetical β-arcade charges of the same kind are very closely stacked on top of each other. This structure is expected to be extremely unstable due to electrostatic repulsion between the buried charged residues. The energy obtained for this structure using the Modeller and FoldX programs was similar to a structure containing polar residues in the core hydrophobic region. This second case is not ideal for fibril formation but is tolerated in amyloids. These results do not reflect reality.

A similar situation occurred with prolines inserted into the β-strand region of fibrils. These insertions are not observed in known β-structures because they should disrupt the β-sheets. However, when the energy obtained for them was compared to corresponding structures without prolines, no significant difference in energy was seen.

The direct, exhaustive energy calculation of all possible β-arcades is another problem – a high number of possible structures needed to be analyzed. Even with a relatively small β-arch of 15 residues, testing all possible sequences of this β-arch requires building and evaluating about 20^{15} (32768000000000000000) structures. This is impossible, as it would take 1039066463723 years if we spend 1s on each. If the fact that a residue may have several rotamers is taken into account the problem is even bigger.

Given these problems it appears that direct application of existing energy calculation programs is not the best way and it is necessary to find another method to overcome all the issues mentioned above. To address this ArchCandy uses empirical rules that first, focus on penalties that allow highly improbable structures to be discarded (Exclusion Rules). Then it scores the remaining structures in a very permissive way by taking into consideration only apparent effects (Scoring Rules). Typically, this prediction yields several possible structures for a given amyloidogenic sequence. The permissiveness of our approach is in agreement with the observed polymorphism of amyloid structures. Indeed, in contrast to globular proteins where one sequence generally corresponds to one 3D structure, one amyloidogenic sequence can have several different amyloid structures (Tycko 2011). The observed polymorphism can be explained by condition-sensitive nucleation sub-structures which can lead to one of several possible structures depending on the conditions.

Thus, until we will know the exact pathways of amyloidogenesis under different conditions, the prediction of the multiple structures will be the only appropriate solution.

3.2.3 The ArchCandy postulated empirical rules

Our general line of reasoning was the following: although we lack complete understanding of the protein structures, thanks to accumulated present day knowledge we have an adequate understanding of the importance of certain major interactions on the stability of the 3D structure. For example, the presence of uncompensated charge residues inside the structure is unacceptable; polar residues with unsatisfied H-bond potential also destabilize the structure, proline breaks β-structural H-bonding, glycines frequently occur in the arc regions, and that the interior of the protein structure is densely packed. Our empirical rules were based on such well established effects. Quantitative estimations and functions used were chosen to fit the results of testing on positive and negative learning subsets composed of long, amyloid forming sequences (Ahmed and Kajava 2013).

ArchCandy analyses protein sequences in three steps: charged residue prefiltering, application of Exclusion Rules and, finally, Scoring.

3.2.3.1 Prefiltering:

Before the detailed analysis of the β-arcade candidates, ArchCandy removes regions that carry an anomalously high charge from the query sequence. The rational behind this filter is that in the parallel in register β-arcade such regions would have very strong electrostatic repulsion. ArchCandy passes a six residue sliding window over the sequence and if the net charge in this window is three or more, the four central residues are removed from further analysis.

3.2.3.2 Exclusion Rules

Steric constraints in the arc region

The steric tension is one of the strongest penalizing interactions. Most of the energy calculation programs correctly evaluate the unfavourable effect of this interaction. Analysing the β-arcade structural models we noticed that some of them, even after a generous energy minimization session continue to have steric tension inside the β-arc regions. This happened when internal residues inside the arc are bulky. In the interior of the arcs, these "tension spots" have either two or three residues in close proximity (Figure 14). This effect, in some cases very severely, limits the amino acid combinations that can be present in this region. The optimal combinations of amino acids are not the same for all β-arches and depend on both the number of residues in the arc and its conformation. To obtain a list of disallowed combinations for each β-arch we undertook their energy evaluation. The basic β-arch used in this analysis had Ala residues in all external positions as the smallest L-amino acid with one rotamer when counting only heavy atoms. The internal positions of β-strands, except the closest to the arc were occupied by Leu residues which provide close packing inside of the β-arcades. For energy minimization and calculation the GROMACS program, Version 3.3.4 (Van Der Spoel *et al.* 2005) was used. The disallowed combinations of the residues in the "tension spots" were then identified by energy calculations. ArchCandy uses this information to remove sub-sequences that contain poor combinations from further analysis. As a result, many combinations that contain bulky aromatic residues were not

allowed. The details for exclusion rules and other rules used by ArchCandy are presented in Annex III.

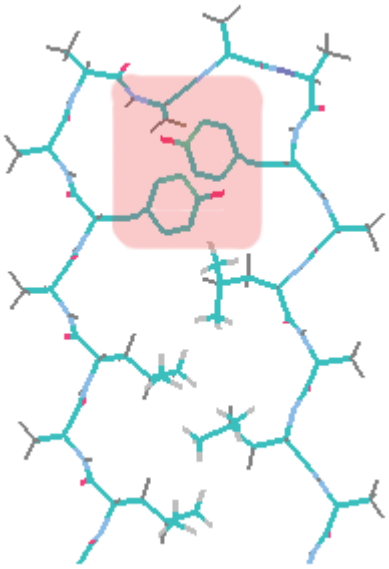


Figure 14. *A "tension spot." The three residues inside the red square are in close proximity. This limits the types of amino acids that can be present in this region. For example, if two of the three residues are bulky residues like Tyr this combination is disallowed since it leads to steric clashes. Visualized using the Coot program (Paul Emsley 2010).*

Charged residues in the β-arcade interior

The charged residues present inside resolved structures of β-arcades are limited to those that are participating in salt-bridges (Luhrs *et al.* 2005; Nelson *et al.* 2005; Iwata *et al.* 2006; Kajava and Steven 2006; Petkova *et al.* 2006; Sawaya *et al.* 2007; Paravastu *et al.* 2008; Qiang *et al.* 2012). A stereo-chemical analysis was used to determine which combinations of the charged residues were capable of forming salt-bridges. Salt bridges are permitted when two side chains with opposite charges were able to reach each other without significant covalent and steric tensions. This test was made by the variation of dihedral angles of the side chains followed by energy minimization. We consider β-arcade structures containing one or more charged residues which are not forming salt bridges inside the structure to be unable to form fibrils.

Prolines in β-strands and arc positions

Prolines in the β-strand region prevent formation of amyloid fibrils since they are unable to take up β-strand conformations and disrupt H-bond network between β-strands. Therefore, we discard all β-arcades that contain Pro in the β-strand regions. They also cannot occur in some positions of the β-arcs which have conformations from the right half of the Ramachandran Plot. These β-arcades were also rejected. To evaluate possibility of prolines to occur at the other positions of the arcs we applied our energy minimization and evaluation procedure. As a result some β-arcades were discarded.

Glycines in β-strands

We consider that a high number of glycines in the β-strand regions disfavour formation of β-arches due to the inability of glycines to provide sufficient van der Waals interactions between β-strands. The presence of glycines also imparts high flexibility to the β-strand which can deter β-arch folding. Therefore, we removed β-arch candidates containing 3 or more glycines in the 4-residues window within the β-strands.

Excess of charged residues

Parallel and in-register β-arcades whose sequences contain a high proportion of charged residues (independently of their sign) are unlikely to occur naturally. Even if the charges are located outside the structure or form salt bridges in the core, residues of the same sign are located on top of the other in the parallel and in-register arrangement. The repulsive electrostatic force of each residue is relatively small, but if there are many of such residues this effect will be considerable. Therefore, we discarded the candidates that have more than 40% charged residues.

3.2.3.3 Optional exclusion rules

Disulphide bond analysis

Cysteine residues may form disulphide bonds in oxidising environments. If these bonds are formed they impose constraint on possible β-arch conformation. Therefore, ArchCandy offers on option that discards the β-arches that are incompatible with the formed SS-bonds. The ability of two cysteines to form an intra-arch SS-bond was tested by stereo-chemical analysis of the structural models. Figure 15 shows examples of allowed and forbidden disulphide bonded β-arches.

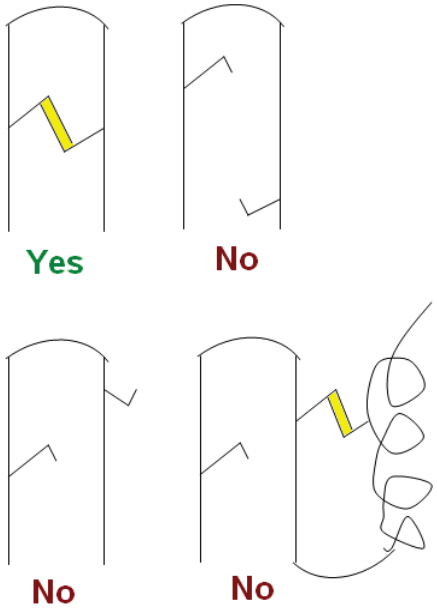


Figure 15. *Allowed β-arches when Disulphide bond analysis is switched on.*

Cavity analysis

Successive small residues facing each other in the β-strand region can lead to the formation of a "cavity" in the otherwise closely-packed structure. This can prevents fibril formation as it disrupts the energetically favourable dense packing of the core. Therefore, ArchCandy has the option that removes β-arcades with such apparent cavities.

Exclusion of Putative Transmembrane Regions

In principle, subsequences of more than 20 residues with high proportion of apolar residues have a high amyloidogenic potential. However, *in vivo* such subsequences can be hidden in the membranes in α-helical conformation. This may prevent formation of the amyloid fibrils by these regions. Therefore, we introduced a filter that can exclude regions predicted to be transmembrane from further analysis.

3.2.3.4 Scoring rules

The Exclusion Rules select the "allowed" candidates, but they are unable to state which of them are more likely to occur. For this purpose ArchCandy uses its Scoring module. The total score of each candidate is a product of eight specific scores (see below). The total and individual scores have values more than 0 and up to 1.

Total Score = β-Strand Length Score* Glycine in Arc Score* Internal AA Composition Score* Total-Net-Charge Score* Charge per Residue Score* Internal Salt-Bridge Score* Arc Length Score* Arc Steric Tension Score

Scores that reflect individual properties are explained in general here. For more details see Annex III.

Arc Steric Tension Score

Exclusion Rules divide the candidates evaluated for steric tension inside their arcs into "disallowed" and "allowed". In accordance with the energy calculations, however, some of the "allowed" candidates still have steric tension. Arc steric tension score introduce penalties for these candidates.

Arc Length Score

If we consider β-strands to be the major structural element that stabilises inter-β-arch interaction, shorter the arc regions lead to smaller entropic loss upon β-arch association making them more favourable for the β-arcade formation. The Arc Length Score introduces this effect into our evaluation.

Glycine in Arc Score

Glycine residues frequently occur in the arc regions of the known β-solenoid structures (Hennetin *et al.* 2006). This can be explained by high flexibility of the glycine-containing regions due to the ability of glycine to take up conformations from all four quadrants of the Ramachandran plot. So the presence of glycines facilitates formation of bends in the polypeptide chain. In addition, generally, arcs are sterically tense and glycines can relieve this tension. Therefore, candidates with arcs containing one or more glycines are not penalized and the score for candidates with arcs not containing glycines have a 0.8 reduction.

Internal AA Composition Score

The composition of the residues in the hydrophobic core of the protein structure determines its stability. The Internal AA Composition Score measures the effects of unfavourable amino acid residues inside the β-arcade structure. Various penalties are associated to polar residues Ser, His, Thr, Cys, and the salt bridges of the charged

residues because even though these residues are involved in salt bridges they are not, in general, able to completely satisfy their H-bonding potentials in the core. Penalties are also applied to Ala, Gly as they contribute to poorer packing in the core due to their small size.

Total-Net-Charge Score

We filter out regions of the sequences that have very high net-charges using the corresponding Exclusion Rules. The Total-Net-Charge score penalises the "allowed" candidates for any deviation of their net charge from zero.

Proportion of Charged Residues Score

Candidates with 40% or more of charged residues are excluded from the subsequent analysis using Exclusion Rules. The *Proportion of Charged Residues Score* estimates the electrostatic repulsion in the sequences with less than 40% of charged residues.

Internal-Salt-Bridge Score

Two kinds of salt-bridges can be formed in the hydrophobic region of the β-arcades: the first is composed of charged residues on two different β-strands of a β-arch, and the second is formed between a residue on the β-arc and one of the β-strands or between two residues of the same β-strand. The former type increases the chances of β-arcade formation as it brings the β-strands together in a fashion that promotes formation of the β-arch. This type of salt-bridge is not penalized. The other types of internal salt-bridges do not have such an effect and the Internal Salt-Bridge Score penalizes these candidates.

β-Strand Length Score

There are limits to how short or how long the β-strand region of a fibril can be. H-bonding between β-strands is the major stabilizing force of the fibrils. Therefore, fibrils become more unstable as their β-strands become shorter. They are also constrained on how long they can be. As β-strands become longer, the surface to volume ratio of the β-arcade, if compared to the ratio of more compact structures, for example, the super-pleated β-structure increases too. As a result, more of the β-arcade becomes exposed to the solution thereby becoming less favourable compared to the alternative structure. Furthermore, all β-arcades show a certain degree of twisting. This means that as the β-strands become longer the residues at the termini of each β-arch move further apart

from the corresponding residues in the β-arches above and below them. This prevents the formation of axial hydrogen bonds. Therefore, the β-strand length score has its maximal value of 1.0 at β-strands of 13 residues and is reduced to zero for the length of less than 5 residues and more than 25 residues (See Annex III for more details).

3.2.4 Procedure of sequence scanning in search of β-arch candidates

There are several hurdles to effectively analysing β-arch candidates in a sequence. Simple sliding window approaches that were used by most of the previous programs for prediction of amyloidogenicity are inadequate because the program must take into account that β-arch candidates are of variable length. Then after an "allowed" β-arcade subsequence is found, the program must decide whether increasing or decreasing the length of the β-strand region will lead to a better fibril forming sequence. Finally, due to the polymorphic nature of amyloids, one fibril forming subsequence may be able to form β-arcades with several varying conformations. A prediction program must not only find these variants, but ideally, also be able to rank them by their likelihood to form fibrils. ArchCandy tackles these issues by scanning the sequence by arc-based expanding windows. The central element of each window is a β-arc with a certain conformation (Figure 16). In this manner a subsequence is analysed to make several candidates for fibril formation. These candidates are assigned scores representing their calculated likelihood of occurrence.

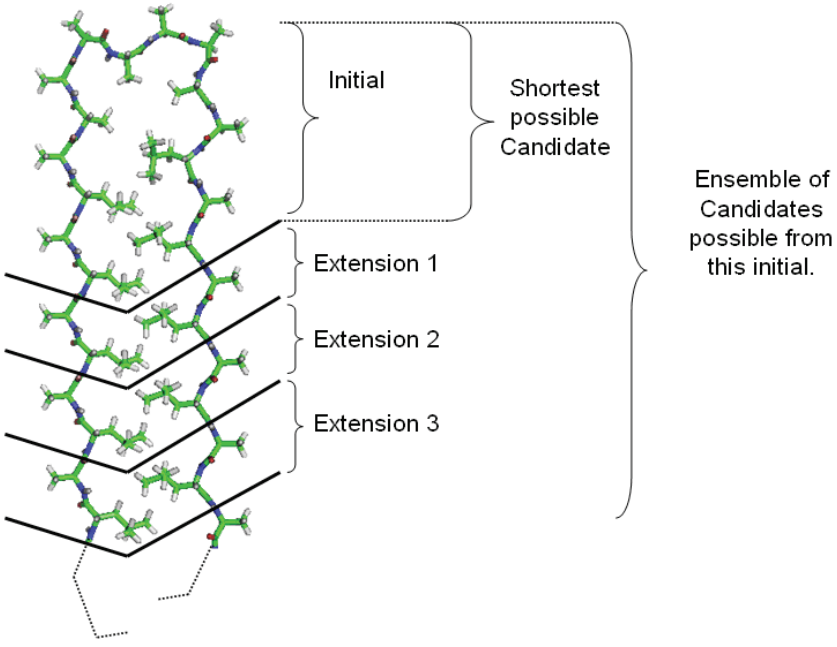


Figure 16. *Composition of **Initials** and **Extensions**.*

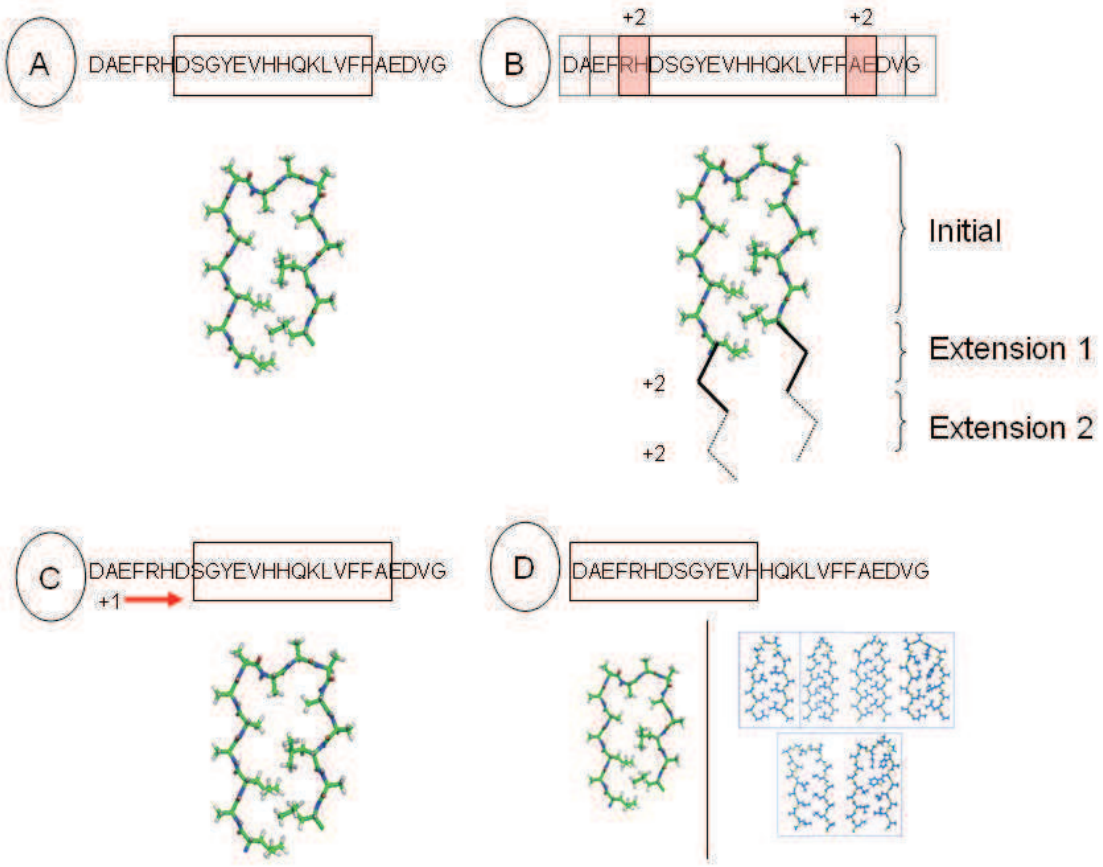


Figure 17. *The arch based expanding window procedure implemented in ArchCandy.*

*A. The sliding window is of the size of an **Initial** (number of residues in the arc region + 10). It represents the smallest possible candidate. Exclusion rules are applied to the **Initial** to determine if it can form a fibril. If it can, the program moves to step B.*

*B. The **Initial** is extended by two residues on either side by expanding the window. Exclusion analysis is applied again. If approved another **Extension** is made. **Extensions** continue to be made until Exclusion analysis fails, the candidate becomes too long for fibril formation, or the end of the sequence is reached.*

*C. When no more **Extensions** can be made they are stored to be scored later. The window then moves forward by one residue. The previous steps are repeated until the end of the sequence.*

*D. The window then returns to the start of the sequence and the **Initials** formed by the 6 other arch templates are analysed one by one.*

The initial window is composed of an arc region flanked by two β-strands of five residues each. This entity is called an **Initial**. Exclusion Analysis is applied to the

Initial, and if approved, this candidate is stored. In the next step, the **Initial** is extended by adding two residues each to the ends of the β-strands (Figure 17). This is called an **Extension**. Exclusion Analysis is then applied to the candidate again. However, this time only the **Extension** is evaluated. If it fails, this candidate is removed from all subsequent analysis. If approved, the candidate is stored and another **Extension** is added and the next candidate is tested. The arc-based candidate will continue to grow for as long as the new **Extensions** are approved, until the candidate becomes too long for fibril formation (more than 50 residues), or if the program reaches the end of the query sequence. All the potential candidates of a given **Initial** are scored using the Scoring module, and the best one is kept for output. Then the same type of the **Initial** moves one position further in the sequence and the procedure of the expanding is repeated. When the scanning is ended by using a given **Initial**, an **Initial** with another β-arc conformation starts to be tested. The query sequence is analysed in this manner seven times, one for each type of the β-arc.

This fashion of scanning the query sequence reflects a probable pathway of the β-arcade nucleation. The nucleation may depend on the inherent ability of protein region to stay some time in short β-arch conformation or β-hairpin (an equivalent of **Initial**) until two such β-arches form a β-arcade nucleus (Kajava *et al.* 2010) and extend their β-strands to the optimal length.

The principle steps of the algorithm can be summarized thusly:

1. Division of query sequence in to arc-based subsequences/candidates.
2. Exclusion Analysis to remove all candidates that cannot form fibrils.
3. Scoring of approved candidates to distinguish those best suited for fibril formation.

3.2.5 ArchCandy Workflow

3.2.5.1 ArchCandy interface for input

The Input interface allows the user to submit the query sequence(s) and choose a scoring threshold and program options (Figure 18). The score threshold can be chosen by the user between 0 and 1. The default value is equal to 0.6 and was established by testing the program against the positive and negative learning datasets (discussed in detail later). Only candidates with scores above the threshold are displayed in the

output. The program has three options that are deactivated by default: “Activate SS-bond analysis”, “Activate Cavity analysis” and “Activate Transmembrane Filter”. The spaces and characters that do not correspond to amino acids are removed from the query sequence prior to the analysis.

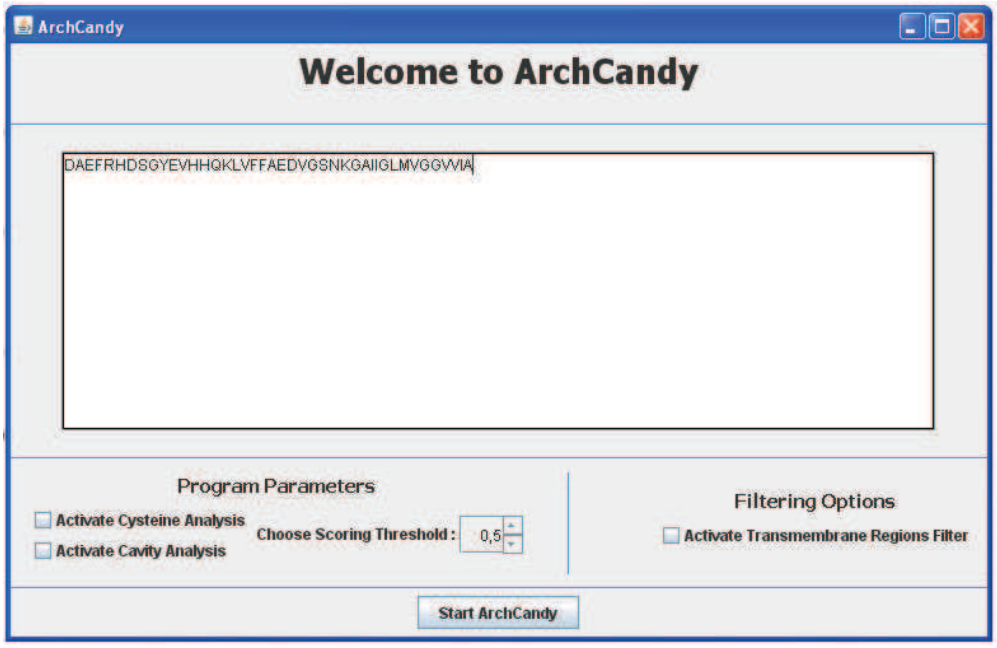


Figure 18. *ArchCandy input interface.*

3.2.5.2 Work of ArchCandy modules for the analysis of the candidates

ArchCandy was written in Java (http://www.java.com/en/). After the input (query sequence) is received the Arch Factory module passes through the query sequence with a sliding window breaking it into numerous β-arc-based candidates. One candidate is created for each iteration of the sliding window. The candidates can be thought of as “virtual β-arches”. Computationally, each candidate is an empty piece of memory with compartments that can only be filled with specific types of information. The first compartment is associated with general information (sequence of the candidate, the header of the query sequence, the type of β-arch template used to analyse the candidate etc). The other two compartments are for information pertaining to the **Initial** and **Extension(s)** respectively. At this point they are empty.

After the candidates are filled with general information by Arch Factory they are sent to the Initial module. Here the **Initials** are tested using exclusion rules and the properties of the **Initial** (results of various tests performed during exclusion analysis, and structural details such as which residues are present in the hydrophobic core) are filled in.

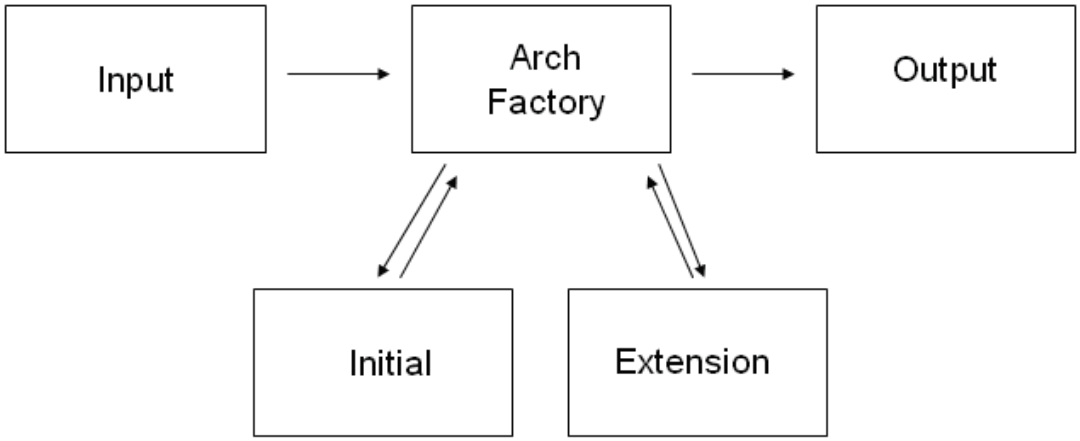


Figure 19. *ArchCandy Workflow.*

All candidates are returned to Arch Factory where those that failed exclusion analysis are removed. The rest are sent to the Extension module. Here the candidates are extended (if possible), evaluated with exclusion analysis, and scored. For each candidate only the **Extension** with the highest score is retained. This data is stored in the extension compartment of the candidate.

The candidates are returned to Arch Factory. This process is repeated for all seven arch templates, creating and filling in several more candidates. Then information from all three compartments of each candidate is used to generate the different kinds of output available in ArchCandy.

This output information is sent to the output module that creates a tabbed window to display the results.

3.2.5.3 ArchCandy interface for output

ArchCandy has several different types of outputs to express the full array of the analyses it conducts. There are five kinds of outputs: **Cumulative histogram, Highest score, SeqView, Table,** and **ScoreCard.**

The **Cumulative histogram** shows the amyloidogenic potential of each amino acid in the sequence. Each bar is the sum of the scores of all candidates that contain that amino acid. It was designed to provide information on the amyloidogenic potential of each residue in the query sequence in a simple visual manner. It can also be used to observe subtle differences in amyloidogenic potential between mutations of the same sequence. Finally, both the candidates with the highest scores and the total number of candidates are both important factors. This is because the tendency of a sequence towards forming many different conformations of amyloids also positively contributes to fibril formation. The cumulative histogram is a means of measuring this effect.

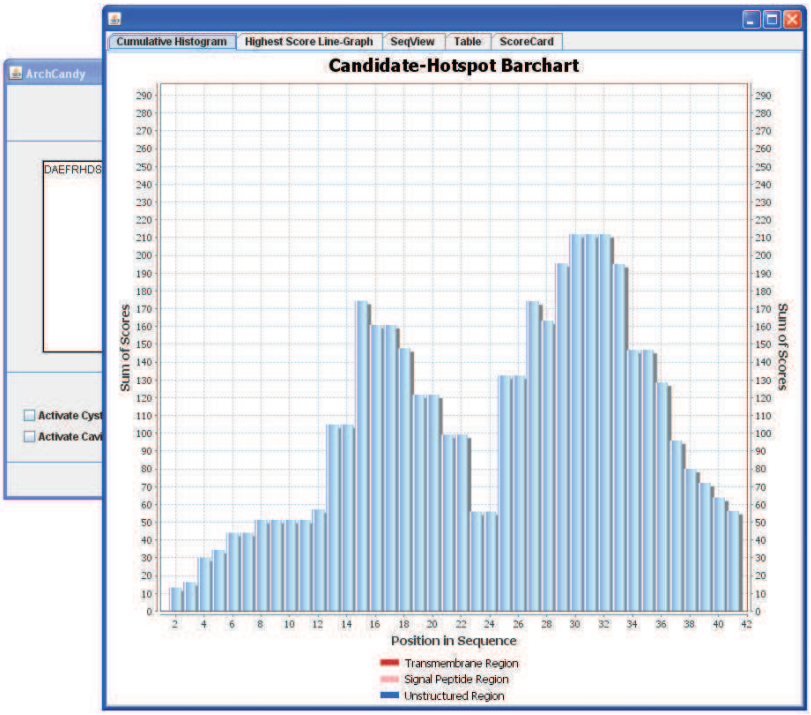


Figure 20. *Cumulative histogram output.*

Highest score is similar to the cumulative histogram. However, in this case the line represents the highest score from all candidates that contain the amino acid. This allows the user to see the regions that are the most amyloidogenic. Cumulative and Highest score histograms are made using the Java library JFreeChart- (http://www.jfree.org/jfreechart/).

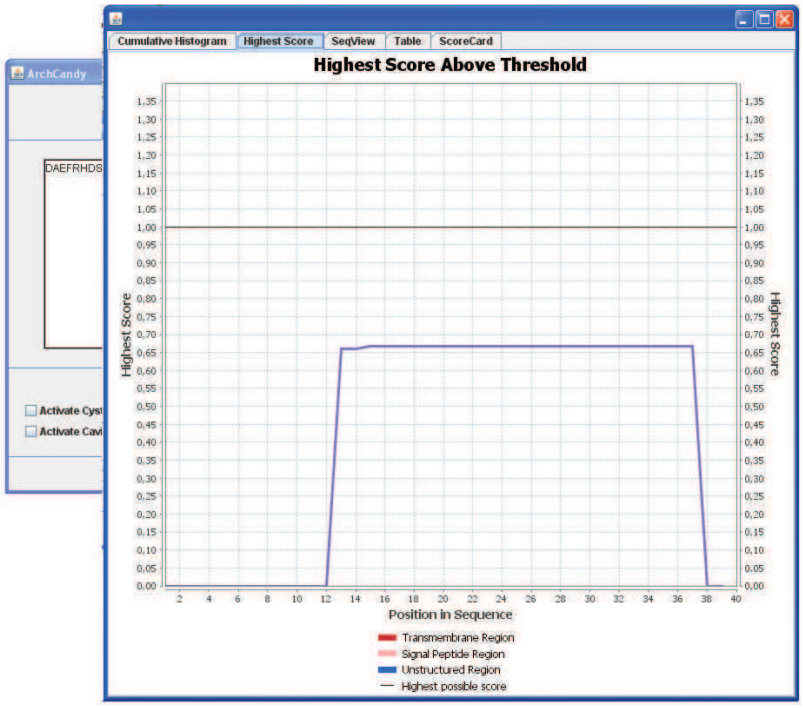


Figure 21. *Highest score output.*

SeqView allows the candidate sequences to be localized and compared to the query sequence. The sequences of all the candidates are aligned to the query sequence and are colour coded with respect to score. The candidates with the highest scores and their positions in the query sequence can easily be determined.

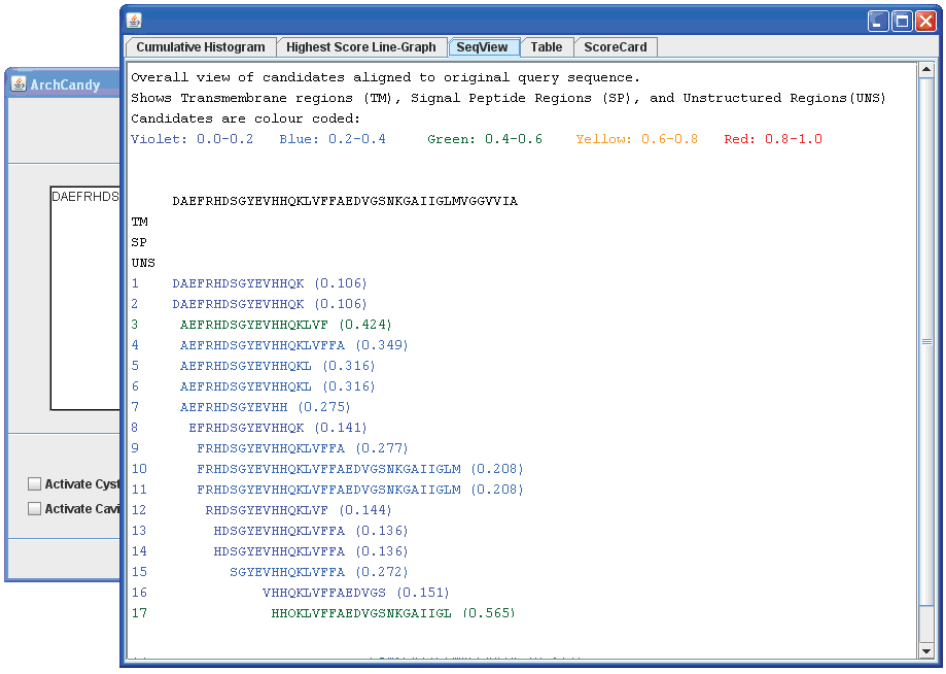


Figure 22. *SeqView output.*

The **Table** shows a 'line-diagram' of the β-arch structure this candidate is predicted to have, its score, the type of arc region it contains, and its position with respect to the query sequence. All the columns can sort from highest to lowest or vice versa when clicked.

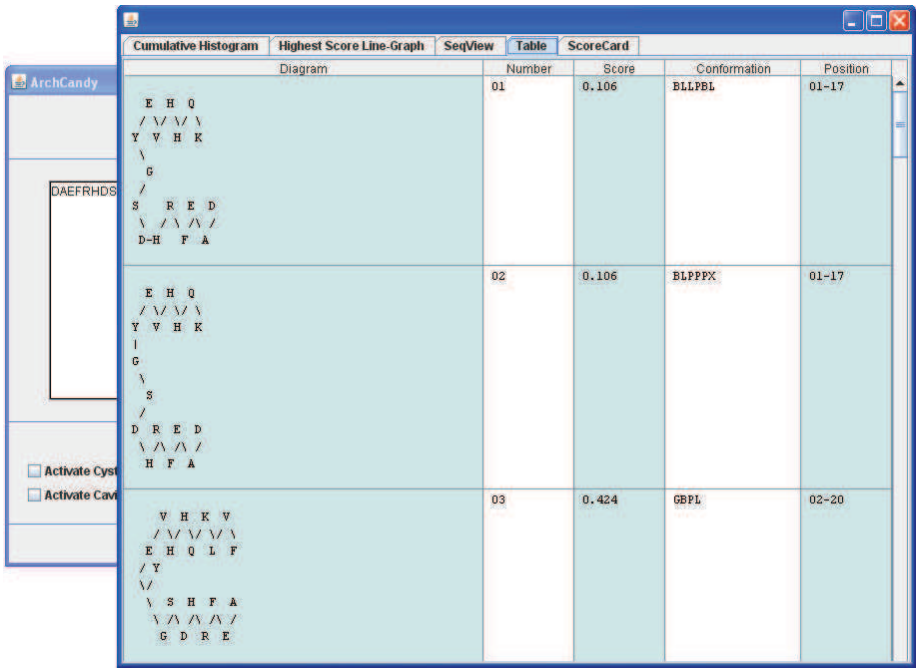


Figure 23. *Table output.*

The **ScoreCard** shows the scoring details of each candidate along with a line-diagram of its predicted structure.

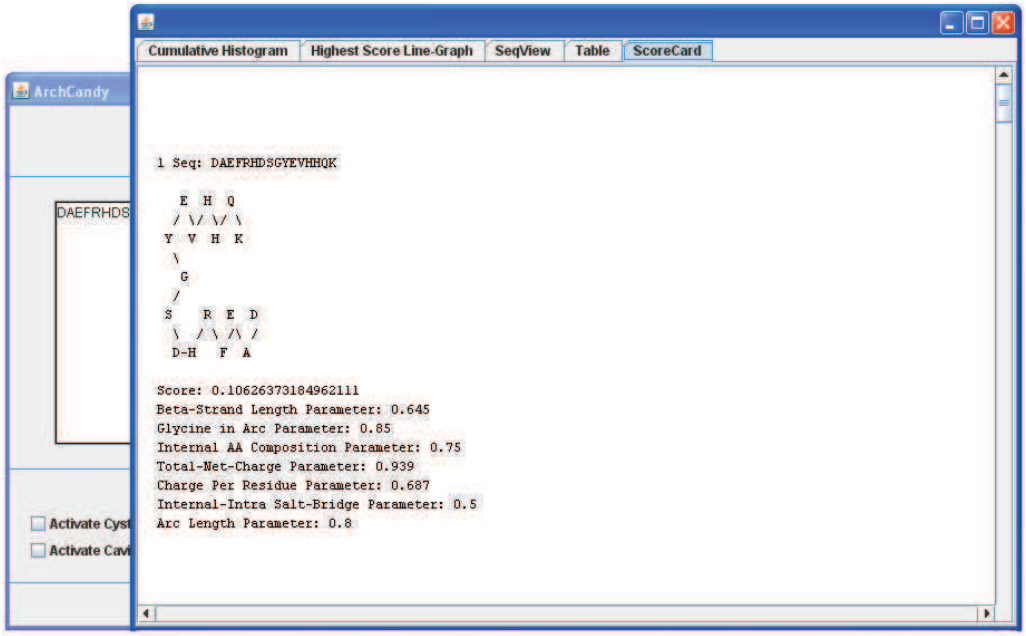


Figure 24. *ScoreCard output*

3.3 Benchmarking ArchCandy

ArchCandy was tested on several datasets to determine different aspects of its predictive ability.

3.3.1 Prediction of Amyloidogenicity

A positive set of 18 amyloid forming proteins and peptides, and a negative set of 52 sequences of non-amyloid-forming and natively unfolded proteins were extracted from literature as described in the previous section (Ahmed and Kajava 2013). During the development of ArchCandy these sets were used to refine the program. Henceforth, they are referred to as the positive and negative learning sets. Care was taken to ensure that sequences were taken from a wide variety of sources to reduce homogeneity. If several similar sequences were found, only one representative sequence was retained in the learning sets. For example, five highly similar Chaplin proteins from *Streptomyces coelicolor* have been shown to form cross-β amyloid fibrils (Claessen *et al.* 2003), but only one of them was used in the positive learning set. Mutants of amyloid-β, amylin and other proteins also were excluded from the positive learning set. This prevents the program from becoming biased towards predicting a certain type of sequences over all others.

After development of ArchCandy was completed, it was tested on positive and negative sets containing all sequences, including mutants found in literature that agreed with our criteria (Annex IV). This is called the extended dataset and contains 52 peptides and proteins in the positive set and 67 in the negative one. Next existing programs were tested on the extended dataset to test their performance in comparison to ArchCandy. The results clearly show superior performance of ArchCandy (Fig. 25).
The Receiver Operator Curve (ROC) was used to establish a score threshold (0.6) for ArchCandy. The threshold represents the best compromise between the highest number of true positives (number of sequences from the positive set that were correctly predicted to be amyloidogenic) and lowest number of false positives (number of sequences from the negative set that were incorrectly predicted to be amyloidogenic).

At score 0.6, ArchCandy correctly predicts 82% of amyloids at very low false positive rate of 0.03%.

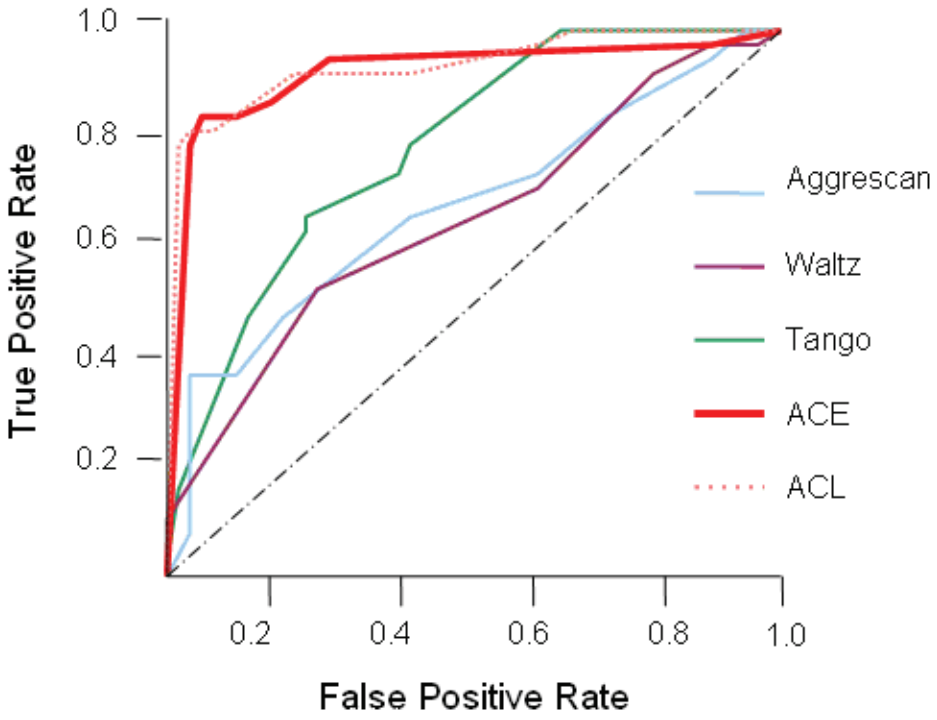


Figure 25. *ROC (Receiver Operator Curve) for Aggrescan, Waltz, Tango, and ArchCandy on the extended dataset (ACE), and ArchCandy on learning dataset (ACL). The programs used in this comparison were chosen for their calculation speed, and ability to process multiple sequences simultaneously. True positive rate= (Number of true positives) / (Total number of amyloid-forming sequences). False positive rate= (Number of false positives) / (Total number of non-amyloidogenic sequences).*

3.3.2 Predicting the Effects of Mutations

It has been shown that protein mutations can increase, decrease, or completely halt the tendency to develop an amyloid (Chiti *et al.* 2003). Several mutant forms of amyloidogenic proteins with increased amyloidogenicity manifest in the human population as familial diseases. Data on the effects of these mutations comes from two sources. Firstly, it is known that a majority of these mutations lead to an early onset of the disease state (for example the Dutch mutation of Amyloid-β) (Zhang-Nunes *et al.* 2006). Secondly, some mutations have also been tested *in vitro* under controlled

conditions to determine their amyloid fibril forming potential. Although it is difficult to compare the effect of mutations described in different publications since a range of conditions have been used and the rate of aggregation can be sensitive to seemingly small changes in buffer or pH, these data are typically used to demonstrate the ability of computer programs to predict the observed change in the amyloidogenicity. A dataset composed of mutants from the human amyloid-β peptide and human amylin was tested with ArchCandy to determine its ability to evaluate the effects of mutations. These two peptides were chosen as they have a large set of the known mutants linked to familial diseases, and because the relatively small size of these peptides ensured more pronounced effects of each single mutation.

Figure 26 shows the results of ArchCandy on various known familial mutations of the amyloid-β peptide.

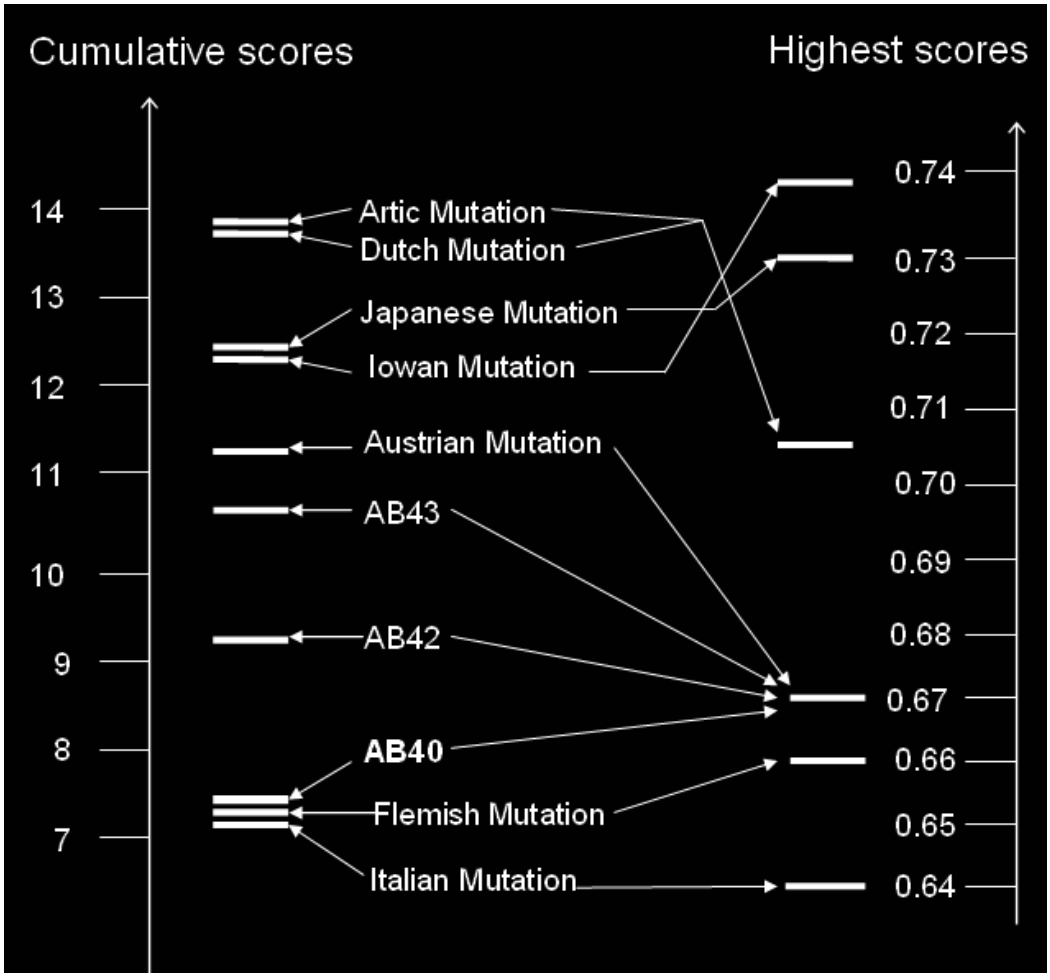


Figure 26. *A scale showing the highest cumulative score (for a given residue a sum of all scores above 0.6) and the highest score associated with various mutations of the amyloid-β peptide. Sequences of the mutants are given in Annex V.*

The results the ArchCandy prediction are in-line with studies conducted on families with individuals manifesting these mutations. In the case of the amyloid-β peptide the fibrillating potential of the mutants is generally compared to their wild type 40 and 42 residue isoforms (Aβ-40 and Aβ-42). Experimentally it was shown that between Aβ-40 and Aβ-42, it is rather Aβ-42 that is linked to the Alzheimer disease (Yin *et al.* 2007). The third isoform, called Aβ-43, is found in the earlier amyloid plagues despite its low level of expression (Parvathy 2001). Thus, the longer is the isoform the larger is its involvement in the Alzheimer disease (Aβ-40 < Aβ-42 < Aβ-43). ArchCandy predicts this tendency (Fig 26). In its turn, the wild type isoforms have lower scores than their mutants known to be involved with more dire disease states (earlier onset, faster progression of disease) (Zhang-Nunes *et al.* 2006) and associated with higher fibrillation potential in accordance with *in vitro* experiments (Nilsberth *et al.* 1999; Miravalle *et al.* 2000; William E. Van Nostrand and Rebeck 2001; Murakami *et al.* 2003; Cloe *et al.* 2011). Exceptions are the Flemish and Italian mutations which have lower ArchCandy scores than the wild type isoforms. The Flemish mutation is associated with an earlier onset of disease (between 35-61 years of age). However, the progression of disease is not completely understood as only two families with a total of 22 infected individuals have been studied (Zhang-Nunes *et al.* 2006). Remarkably, the *in vitro* results for the Flemish mutation show that it forms fibrils just as well or slightly less readily than Aβ-40 (Van Nostrand *et al.* 2001), in agreement with ArchCandy predictions. The Italian mutation is also associated with an earlier onset of disease (between 62-75 years of age). However, mature senile plaques are not found, and the diffuse deposits that are present do not stain with ThT. This suggests that β-sheet rich structures may not be the principle actors involved in the Italian Alzheimer disease (Zhang-Nunes *et al.* 2006). The Italian mutation, however, has been shown to be both equally and more amyloidogenic than Aβ-40 *in vitro* (Miravalle *et al.* 2000; Murakami *et al.* 2003).

It is worth mentioning that at present it is not clear which score is better to use for interpretation of the ArchCandy prediction: cumulative score or the highest score. In the case of the amyloid-β mutants, our prediction results show that the cumulative score agrees better with the observed data than the highest score. When we consider the highest score (shown on Fig. 26 as scale on the right) Aβ-40, Aβ-42, Aβ-43, and the Austrian mutation, all have equivalent highest scores. At the same time, the cumulative

highest score is able to predict the observed differences in amyloidogenic potential of these peptides.

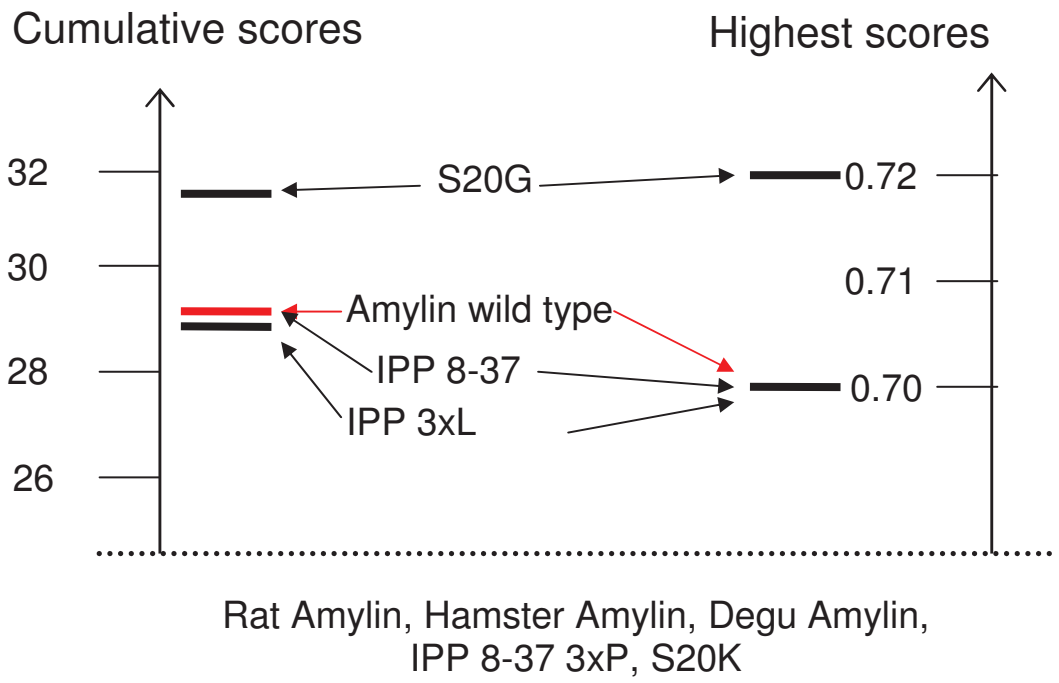


Figure 27. *Human amylin mutants. Cumulative highest score and highest scores for disease related mutations of the human amylin peptide. Mutations under the dotted line are those not predicted to be amyloidogenic. Sequences of the mutants are given in Annex V.*

The same analysis was done for amylin (Fig. 27). The Japanese (S20G) mutation of human amylin is associated with an earlier onset of diabetes mellitus type 2 disease in patients (Sakagashira *et al.* 1996; Sakagashira *et al.* 2000). ArchCandy is able to correctly predict this effect. It has also been established that rat, hamster and degu amylins do not form fibrils (Westermark *et al.* 1992). ArchCandy scores for these peptides are below the 0.6 threshold being in agreement with the experiment. In addition, several mutations of amylin have been tested *in vitro* for fibril formation [S20K, 3xL (F15L/F23L/Y37L), 8-37 3xP (V17P/S19P/T30P) and amylin 8-37] (Abedini and Raleigh 2006; Marek *et al.* 2007; Cao *et al.* 2012). The N-terminally truncated amylin 8-37 has the same fibril forming potential as the wild type (Abedini and Raleigh 2006) which is correctly predicted by ArchCandy. The amylin 3xL mutant has less fibrillation potential than the wild type (Marek *et al.* 2007) and this is reflected

in the cumulative score of ArchCandy. Amylin (8-37) 3xP does not form the fibrils (Abedini and Raleigh 2006) and this result is predicted by ArchCandy. Finally, the S20K mutation has been shown to both have lower amyloidogenicity and to not form fibrils at all (Cao *et al.* 2012). ArchCandy predicts the later.

At the same time, the results of the ArchCandy prediction disagree with the observed effect of a series of mutations S28G, I26D, A13E, L16Q published in one of the publications (Fox *et al.* 2010). In this work, the authors tested fibril-forming potential of amylin peptide fused to GFP. Interpretation of this result need to be taken with precaution, due to the fact that the steric repulsion of the folded GFP structures can prevent formation of the fibrils

3.3.3 Prediction of Localization of Amyloidogenic Regions within Proteins

ArchCandy is able to not only predict the amyloidogenicity of sequences, but also to correctly identify the regions within the sequence that form fibrils. It is demonstrated on several large proteins with the known location of the amyloidogenic regions such as Sup35p, Ure2p, Receptor-interacting serine/threonine-protein kinase 3 (RIP3), and TAR DNA-binding Protein (TDP) (Baxa *et al.* 2006; Chen *et al.* 2010; Li *et al.* 2012). Tests of these proteins show that ArchCandy has an inherent advantage over existing programs. It correctly assign the highest cumulative scores to experimentally identified amyloidogenic regions and predicts low scores for the remaining part of the proteins independently of whether these parts are naturally unfolded or have globular structures (Fig. 28-31). In contrast, the other existing programs have a tendency to predict amyloidogenic regions all over the sequences.

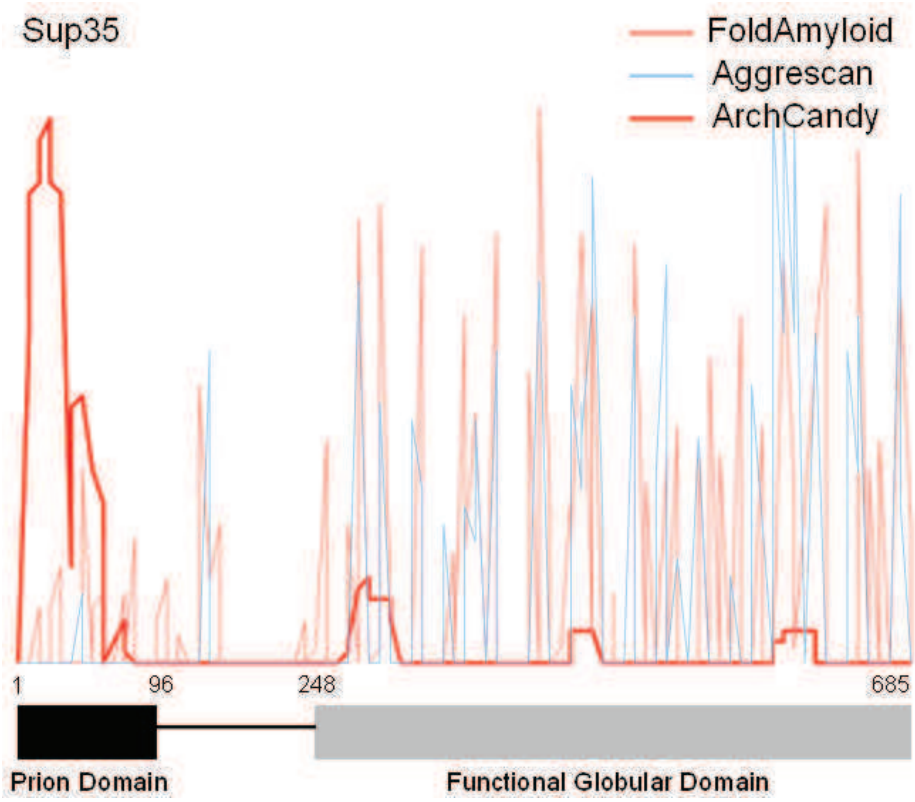


Figure 28. *Localization results for FoldAmyloid, Aggrescan, and ArchCandy on Saccharomyces cerevisiae Sup35. Location of prion domain as determined by (Baxa et al. 2006).* Here and in Figures 29, 30 and 31 *underneath the graph the black block represents the prion domain or amyloidogenic region, the grey block shows the functional globular domain, and the connecting line corresponds to unfolded regions. For FoldAmyloid and Aggrescan all negative values were changed to zero. The Waltz program (Maurer-Stroh et al. 2010) gives results similar to Tango, FoldAmyloid, and Aggrescan; however, it provides only a graph to show localization, without values for the amyloidogenic potential of each residue. Therefore, it is not present here. For sup35p Tango results are not present as it is unable to handle a sequence of this length.*

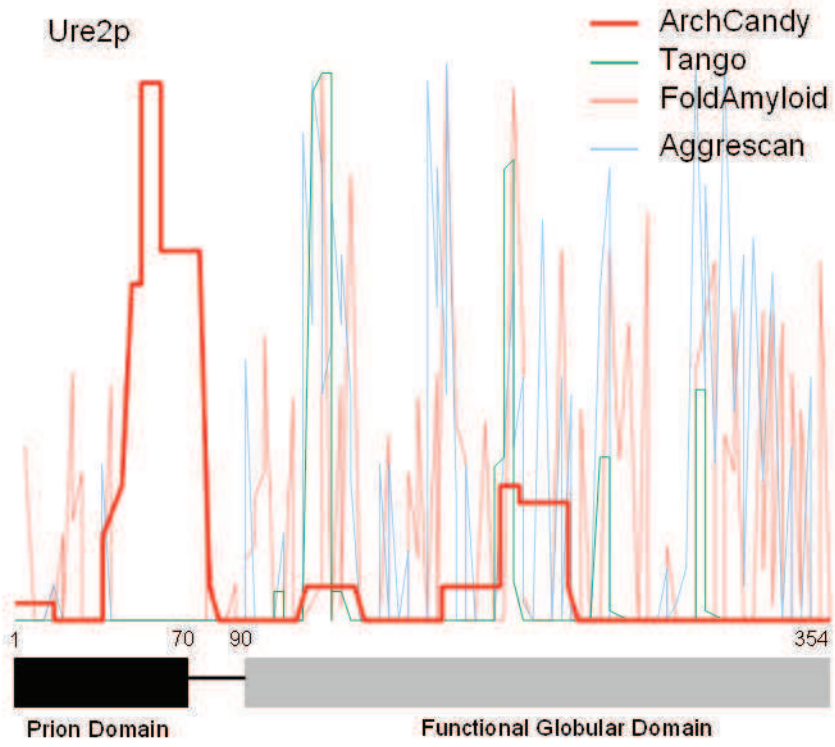


Figure 29. *Localization results for Tango, FoldAmyloid, Aggrescan, and ArchCandy on Saccharomyces cerevisiae Ure2p. Localization for prion domain as determined by (Baxa et al. 2006).*

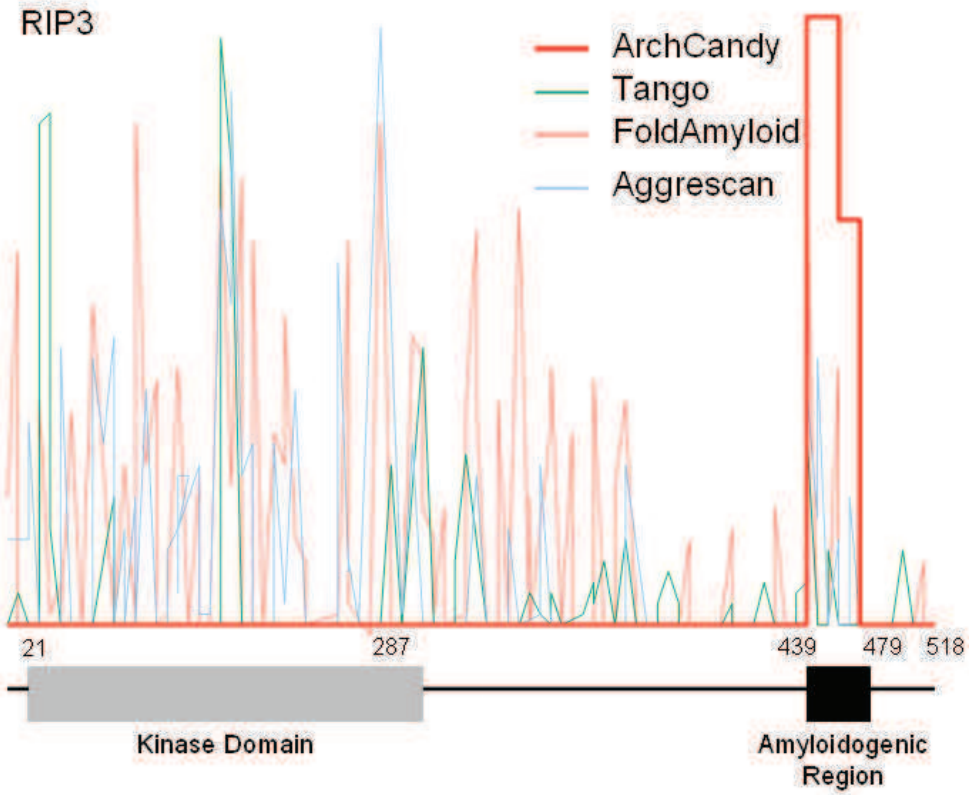


Figure 29. *Localization results for Tango, FoldAmyloid, Aggrescan, and ArchCandy on human RIP3. Amyloidogenic region as determined by (Li et al. 2012)*

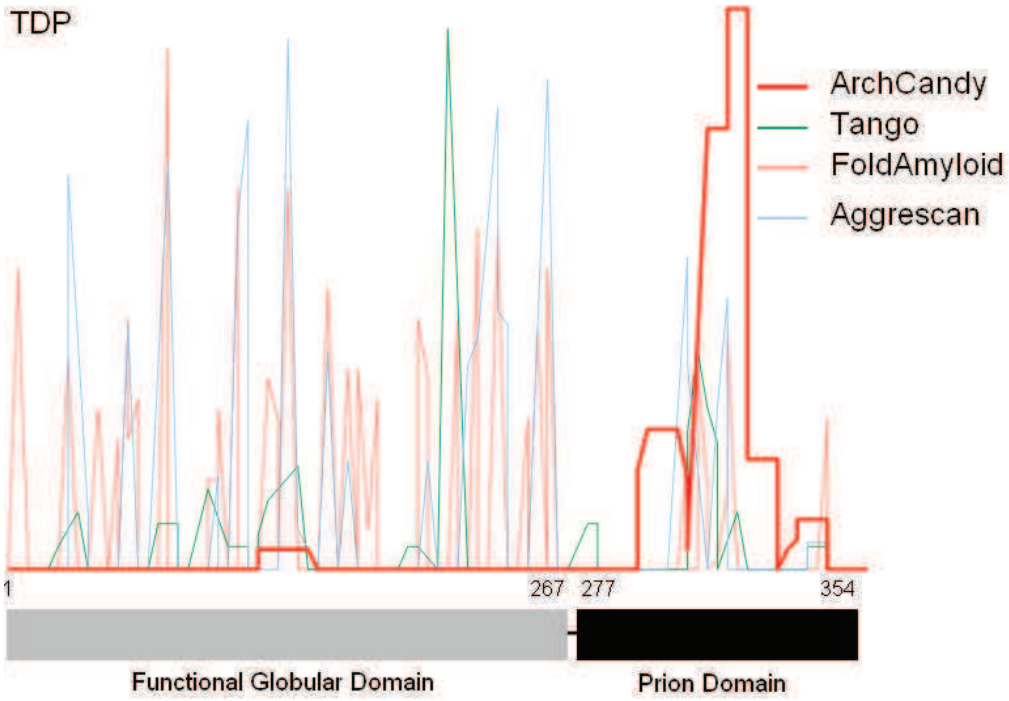


Figure 30. *Localization results for Tango, FoldAmyloid, Aggrescan, and ArchCandy on human TDP. Location of amyloidogenic region as determined by (Chen et al. 2010).*

It is worth mentioning that the size of the long (over 40 residues) amyloidogenic regions is frequently overestimated and the exact boundaries of these regions within proteins remain to be determined. For example, Sup35 prion domain is assign to the first 90 residues of the protein; however, there are a number of data showing that the prion domain is shorter (Osherovich *et al.* 2004; Nelson *et al.* 2005). In this situation, ArchCandy prediction of β-arches may be used to guide mutational analysis to better establish amyloidogenic regions.

3.3.4 Prediction of 3D structure of β-arcades

ArchCandy was conceived in a manner that allows prediction of the conformation attained by predicted fibrils. These predictions can be seen in Table view of each candidate (Figure 23). To check the accuracy of these predictions they were compared to the experimentally resolved structures of proteins and peptides.

There are two resolved 3D structures of amyloid-β which correspond to two different types of amyloid fibrils (Luhrs *et al.* 2005; Petkova *et al.* 2006). Among the 4 β-arch candidates (with score above 0.6) proposed by ArchCandy there is one (with the 3rd highest score) that exactly corresponds to the 3D structure of an amyloid-β fibrils (Luhrs *et al.* 2005) (Figure 31).

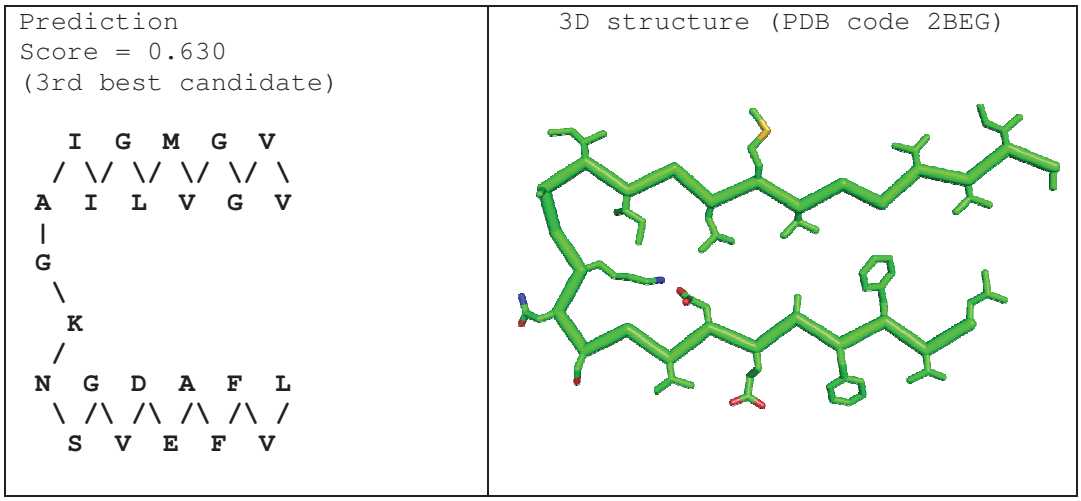


Figure 31. *2-D diagram of β-arches predicted by ArchCandy for amyloid-B (left), and the 3D-structure of β-arch resolved by ssNMR, (right). PDB code: 2BEG (Luhrs et al. 2005).*

ArchCandy also correctly predicts the β-structural arrangement of the second amyloid-β fibril (Petkova *et al.* 2006). Concerning β-arcs in this 3D structure, each β-arch has a different arc conformation within the stack of several β-arches. Among these β-arches there is one that perfectly agrees with ArchCandy prediction (Figure 32).

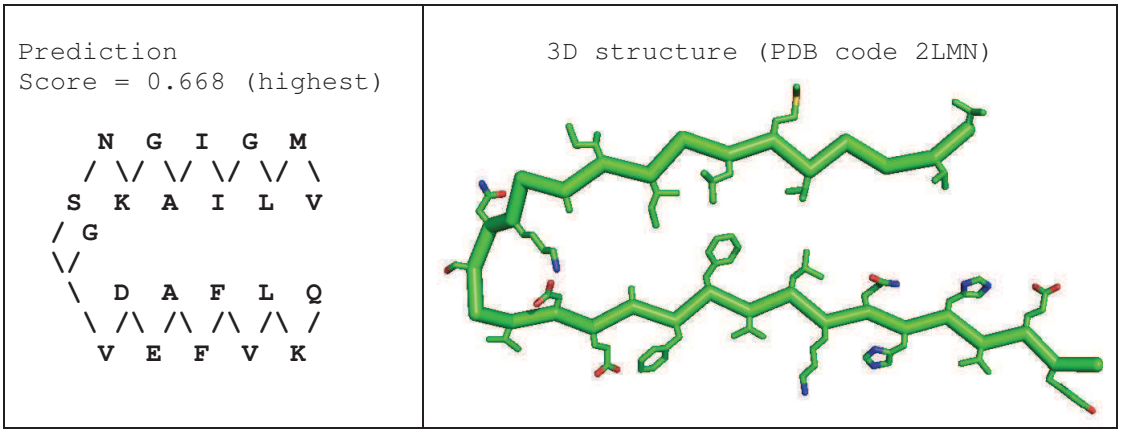


Figure 32. *2-D diagram of β-arches predicted by ArchCandy for amyloid-β (left), and the 3D-structure of β-arch resolved by ssNMR, (right). PDB name: 2LMN (Petkova et al. 2006).*

The 3D structure of the Iowan mutant of amyloid-β has also been determined (Qiang *et al.* 2012). ArchCandy correctly predicts its β-arch (Figure 33). At the same time, the resolved fibrils are formed by anti-parallel arrangements of such β-arches and this successful prediction can be considered rather as a co-lateral success, because ArchCandy is tuned to predict parallel and in-register β-arcades.

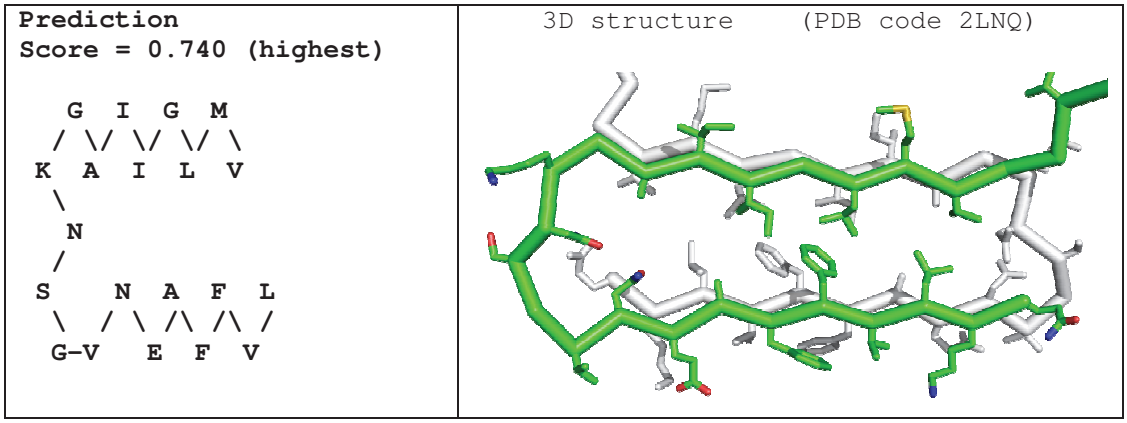


Figure 33. *2-D diagram of β-arches predicted by ArchCandy for amyloid-β (left), and the 3D-structure of β-arch of anti-parallel β-sheet architecture in Iowa-mutant amyloid-β fibrils (right). PDB code: 2LNQ (Qiang et al. 2012).*

The other correct prediction is related to the fibrils formed by human CA150 protein, a transcriptional activator that binds to and is co-deposited with huntingtin during Huntington's disease (Ferguson *et al.* 2006) (Figure 34). Interestingly, it is known that a mutant Arg24Ala of CA150 does not form fibrils (Ferguson *et al.* 2006) and ArchCandy predicts this effect.

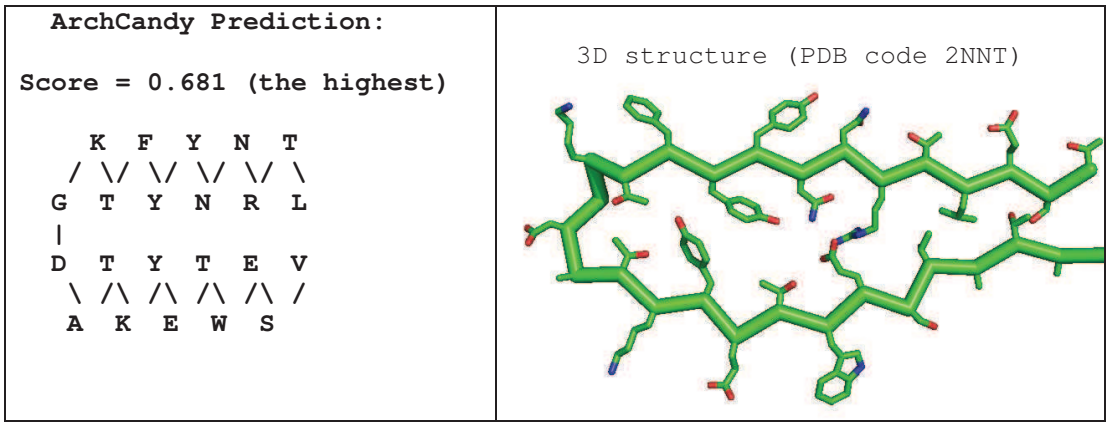


Figure 34. *2-D Diagram of β-arches predicted by ArchCandy for Human CA150 (left), and the 3D-structure of Human CA150, (right). PDB code: 2NNT. (Ferguson et al. 2006).*

Finally, the fifth resolved structure is a protofilament of β2-microglobulin fragment (Iwata *et al.* 2006). One *β*-strand of this structure has an excess of negatively charged residues (Fig. 35).

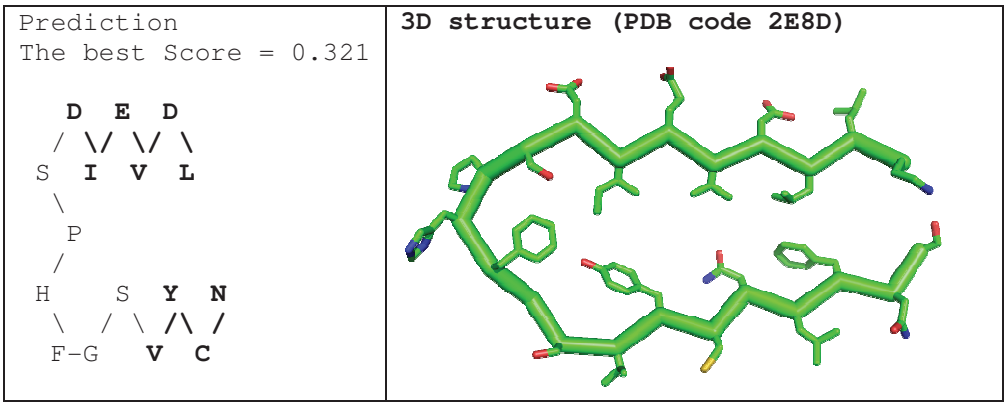


Figure 35. *2-D Diagram of β-arches predicted by ArchCandy (left), and the 3D structure of amyloid protofilaments of beta2-microglobulin fragment probed by solid-state NMR. PDB code: 2E8D (Iwata et al. 2006).*

Therefore, ArchCandy scores this peptide below the threshold of 0.6. At the same time, its *β*-arch candidate with the highest score of 0.321 corresponds to the correct β-strand arrangement (Figure 35). The prediction, however, differs in the arc region. In a situation similar to one of the amyloid-*β* fibrils (Petkova *et al.* 2006), the *β*-arcade of β2-microglobulin fragment has different arc conformations in each β-arch. Such a strong variation of the arc conformation suggests that these structures are not well-resolved and may contain mistakes. Indeed, it is known that identical blocks of high resolution crystal structures have the same conformations (in the approximation that does not take into account some variation of the side-chain rotamers) (Kajava 2012). Taking this fact into consideration, we conclude that the structures mentioned above (Iwata *et al.* 2006; Petkova *et al.* 2006) may require the refinement. In fact, the ssNMR structures are usually obtained by MD simulations of the constrained models. This procedure may be a source of erroneous conformations. In this situation, arcs conformations suggested by ArchCandy (they were chosen from the frequently occurring arcs of the crystal structures (Hennetin *et al.* 2006)) can be used for the refinement of the ssNMR structures. For example, we suggest that amyloid-*β* fibril structure studied by (Petkova *et al.* 2006) consist of one type of *β*-arches with "gbpl" conformation of arcs that is shown on Figure 32.

Although ArchCandy was designed to predict β-arches of parallel and in-register β-arcades, it also correctly predicted the β-arch arrangement within the anti-parallel structure of Iowan mutant of amyloid-β (Qiang *et al.* 2012) (Figure 34). Furthermore, tests of ArchCandy against proteins that are known to form cross-β amyloids from stacks of β-solenoids (Het-s prion (Wasmer *et al.* 2008) and CsgA amyloids(Wang *et al.* 2005)) reveals that their scores are also high (close or above 0.6). This suggests that ArchCandy can be also used for prediction of β-arches in the other β-arch-containing fibrils such as with anti-parallel structure or with stacks of β-solenoids.

4. Discussion and Perspectives

Current programs for amyloid prediction are unable to make use of the full ensemble of recently obtained structural information. The objective of this work was to fill this void and to develop a new approach based on the assumption that sequences that are able to form β-arcades are amyloidogenic. We have described the development of the algorithm and a computer program called ArchCandy. The results obtained with ArchCandy on a wide variety of datasets have shown that it performs better than previously existing programs. ArchCandy is able to distinguish between longer, naturally occurring, disease related amyloidogenic and non-fibril forming sequences, to explain the effect of mutations on the fibril forming potential of proteins, it has been shown to localize known amyloidogenic regions correctly, and it can predict the 3D structures of the β-arches of fibrils.

However, ArchCandy has certain limitations. For example, by default ArchCandy considers all predicted structures to be composed of in-register parallel β-arches. Indeed, this type of amyloid fibrils is the most frequent. However, it is known that fibrils can be formed by the stacking of anti-parallel β-arches (Qiang *et al.* 2012), or β-solenoidal structures (Wang *et al.* 2005; Wasmer *et al.* 2008). Furthermore, amyloidogenic regions that are longer than one β-arch can form superpleated β-structures that consist of several β-arches concatenated into serpentines (Kajava *et al.* 2004). The current version of ArchCandy is not designed to predict other β-arch arrangements. However, efforts will be made to incorporate a module for assessment of these β-arch structures. A few fibrils are also formed by the interaction of globular domains to each other (Nelson and Eisenberg 2006; Chiti and Dobson 2009). However, they are out of the scope of this work.

The majority of protein sequences that form amyloid fibrils are unfolded in their native state. Folded polypeptide chains may also contain amyloidogenic regions within them. However, as these amyloidogenic regions are hidden within the 3D structure, they are not available for fibril-formation. Significant efforts have been dedicated to the identification of such hidden regions (also known as ‘conformational switches’ or “chameleon” sequences) within globular proteins that are innocuous in their normal

state (Chiti *et al.* 2000; Yoon and Welsh 2004; Tartaglia and Vendruscolo 2008; Kim *et al.* 2009). ArchCandy partially takes these effects into account using the "Optional Exlusion Rules," excluding transmembrane regions and β-arches that are incompatible with known disulphide bonds. Although efforts were not made during ArchCandy development to tackle this problem directly, this version of the program was surprisingly able to distinguish between amyloidogenic regions in proteins containing functioning globular domains quite well (Figures 28, 29, 30, 31). Special efforts are planned to improve this aspect of the ArchCandy.

ArchCandy was developed for typical physiological conditions and in particular for range of pH (6-8) when Asp/Glu and Lys/Arg are negatively and positively charged correspondingly. However, one may be interested to test amyloidogenicity of peptides or proteins in acid or basic pH. The fact that the charged side-chains can become neutral at certain pH is not accounted for in ArchCandy. However, the user can approximate this phenomenon by changing in the input file the negatively charged residues Glu and Asp to Gln and Asn below their pKa values, and Lys and Arg above their pKa values to a neutral residue (for example, His). The post-translational modifications such as phosphorylation can be taken into account in a similar manner by substitution in the imput sequence a phosphorylated residue to Glu.

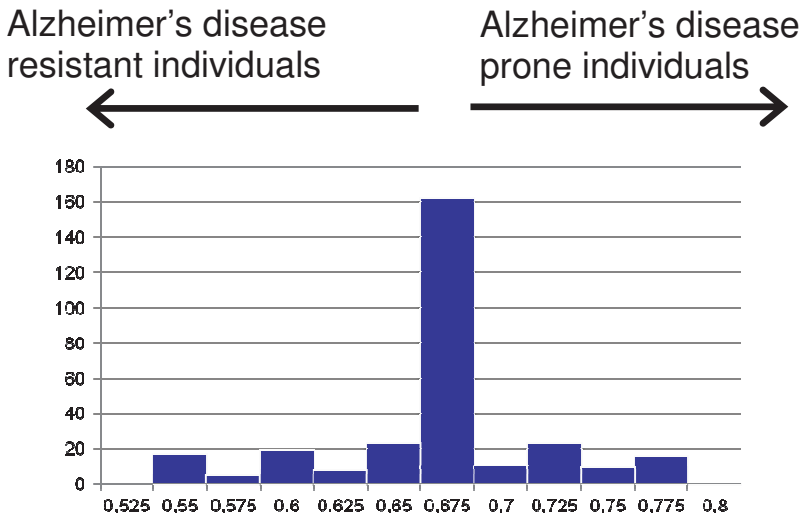
5. Conclusions

Numerous studies have shown that the ability to form amyloid fibrils is an inherent property of the polypeptide chain. This has lead to the development of a number of computational approaches to predict amyloidogenicity by amino acid sequences. However, existing methods generate an unsatisfactorily high number of false positives when tested against longer sequences of the disease-related peptides and proteins. In this work we developed an improved bioinformatics based approach to predict amyloidogenic regions from protein sequences.

Our results show a high level of performance in the prediction of amyloidogenic regions. In addition to the purely academic significance of the results achieved ArchCandy opens exiting avenues for several important applications in biotechnology, the pharmaceutical industry, and medicine. Aggregation is often a bottleneck in the production of recombinant proteins. ArchCandy can potentially address this problem with its ability to detect the amyloidogenic regions and suggest mutations that will make aggregation prone proteins soluble.

Since ArchCandy also predicts the atomic structure of the β-arcades, it can be used in combination with the experimental data for the refinement of 3D structures. In the results section we describe the cases of structures proposed by (Iwata *et al.* 2006; Petkova *et al.* 2006) that may benefit from this approach. This can potentially be used to obtain more precise structures of amyloid fibrils which will allow structure-based drug design protocols in search of inhibitors of amyloidosis.

Finally, amyloid prediction tools are particularly relevant to the disease-related amyloids as currently no reliable ways to diagnose the early stages of such diseases are available. Thanks to a radical drop in the cost of sequencing an individual's genome, such bioinformatics tools are becoming extremely timely. With further research, an accurate risk profile might enable individuals to take steps to prevent diseases for which they are at increased risk based on genetics.



>Human amyloid-β42
DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA

Figure 36. *ArchCandy analysis of 300 single point, random mutations of the amyloid-β peptide made using the Artificial Life Framework program (Dalquen et al. 2012).*

Generally, the disease related mutations are known, as they are the ones that have been studied. However, there is a strong possibility that protective mutations also occur undetected in the human population. ArchCandy provides an avenue for the prediction of the effects of these mutations. Our preliminary test of ArchCandy conducted on 300, computationally produced, random mutations made to the amyloid-β peptide shows an almost equal mix of protective and amyloid forming mutants (Figure 36). Finally, ArchCandy can potentially be used in the large scale analysis of proteomes to find new amyloidogenic proteins.

Annex I

List of Abbreviations

aapv : average aggregation propensity value

ACE : ArchCandy on the Extended dataset

ACL : ArchCandy on Learning dataset

AFM : Atomic Force Microscopy

Aβ-42 : 42 amino acid human peptide β-amyloid

BSE : Bovine Spongiform Encephalopathy

CPEB : Cytoplasmic Polyadenylation Element Binding

EM : Electron Microscopy

GFP : Green Fluorescent Protein

GROMACS : GROningen Machine for Chemical Simulations

H : hydrogen

HST : hot spot threshold

IB : Inclusion Bodies

NMR : Nuclear Magnetic Resonance

PSSM : Position Specific Scoring Matrix

RIP3 : Receptor-Interacting serine/threonine-Protein kinase 3

TDP : TAR DNA-binding Protein

ROC : Receiver Operator Curve

SP : Start Position

ssNMR : solid state NMR

ThT : Thioflavin T

Three letter codes for all natural amino acids

Annex II

Negative Dataset for Testing Amyloid Prediction Programs

The negative set was extracted from the DisProt database of disordered proteins (Vucetic *et al.* 2005) with the following criteria: sequences are disordered in their entirety and have less than 150 residues.

The negative set contains 52 sequences.

>DisProt|DP00001|uniprot|Q9HFQ6|sp|RLA3_CANAL #1-108
MSTEASVSYAALILADAEQEITSEKLLAITKAAGANVDQVWADVFAKAVEGKNLKELLFSFAAA
APASGAAAGSASGAAAGGEAAAEEAAEEEAAEESDDDMGFGLFD
>DisProt|DP00002|uniprot|P02400|sp|RLA4_YEAST #1-110
MKYLAAYLLLVQGGNAAPSAADIKAVVESVGAEVDEARINELLSSLEGKGSLEEIIAEGQKKFAT
VPTGGASSAAAGAAGAAAGGDAAEEEKEEEAKEESDDDMGFGLFD
>DisProt|DP00004_C002|uniprot|P49913|unigene|Hs.51120|sp|CAMP_HUMAN #1-37
LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
>DisProt|DP00005|uniprot|P03045|sp|REGN_LAMBD #1-107
MDAQTRRRERRAEKQAQWKAANPLLVGVSAKPVNLPILSLNRKPKSRVESALNPIDLTVLAEYH
KQIESNLQRIERKNQRTWYSKPGERGITCSGRQKIKGKSIPLI
>DisProt|DP00006|uniprot|P00004|unigene|Eca.1571|sp|CYC_HORSE #1-104
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEET
LMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
>DisProt|DP00022|uniprot|P17639|sp|EMB1_DAUCA #1-92
MASQQEKKELDARARQGETVVPGGTGGKSLEAQQHLAEGRSKGGQTRKEQLGGEGYHEMGRK
GGLSNNDMSGGERAEQEGIDIDESKFRTKK
>DisProt|DP00024|uniprot|P03129|sp|VE7_HPV16 #1-98
MHGDTPTLHEYMLDLQPETTDLYCYEQLSDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCKCDSTL
RLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP
>DisProt|DP00027|uniprot|P26477|sp|FLGM_SALTY #1-97
MSIDRTSPLKPVSTVQTRETSDTPVQKTRQEKTSAATSASVTLSDAQAKLMQPGVSDINMERVEA
LKTAIRNGELKMDTGKIADSLIREAQSYLQSK
>DisProt|DP00028|uniprot|Q13541|unigene|Hs.411641|sp|4EBP1_HUMAN #1-118
MSGGSSCSQTPSRAIPATRRVVLGDGVQLPPGDYSTTPGGTLFSTTPGGTRIIYDRKFLMECRNSP
VTKTPPRDLPTIPGVTSPSSDEPPMEASQSHLRNSPEDKRAGGEESQFEMDI
>DisProt|DP00039|uniprot|P05204|unigene|Hs.181163|sp|HMGN2_HUMAN #1-89
PKRKAEGDAKGDKAKVKDEPQRRSARLSAKPAPPKPEPKPKKAPAKKGEKVPKGKKGKADAG
KEGNNPAENGDAKTDQAQKAEGAGDAK

>DisProt|DP00040|uniprot|P17096|unigene|Hs.518805|sp|HMGA1_HUMAN #1-107
MSESSSKSSQPLASKQEKDGTEKRGRGRPRKQPPVSPGTALVGSQKEPSEVPTPKRPRGRPKGSK
NKGAAKTRKTTTTPGRKPRGRPKKLEKEEEEGISQESSEEEQ
>DisProt|DP00057|uniprot|P15340|sp|HSP1_CHICK #1-62
MARYRRSRTRSRSPRSRRRRRRSGRRRSPRRRRRYGSARRSRRSVGGRRRRYGSRRRRRRRY
>DisProt|DP00058|uniprot|P06302|unigene|Rn.817|sp|PTMA_RAT #1-112
MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAQNEENGEQEADNEVDEEEEEGGEE
EEEEEEGDGEEEDGDEDEEAEAPTGKRVAEDDEDDDVETKKQKKTDEDD
>DisProt|DP00070|uniprot|P37840-1|unigene|Hs.21374|sp|SYUA_HUMAN #1-140
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKE
QVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA
>DisProt|DP00116|uniprot|P81455|sp|OSTCN_CANFA #1-49
YLDSGLGAPVPYPDPLEPKREVCELNPNCDELADHIGFQEAYQRFYGPV
>DisProt|DP00140|uniprot|P0A7L8|sp|RL27_ECOLI #1-85
MAHKKAGGSTRNGRDSEAKRLGVKRFGGESVLAGSIIVRQRGTKFHAGANVGCGRDHTLFAKA
DGKVKFEVKGPKNRKFISIEAE
>DisProt|DP00143|uniprot|P0A7N9|sp|RL33_ECOLI #1-55
MAKGIREKIKLVSSAGTGHFYTTTKNKRTKPEKLELKKFDPVVRQHVIYKEAKIK
>DisProt|DP00145|uniprot|P0A7S3|sp|RS12_ECOLI #1-124
MATVNQLVRKPRARKVAKSNVPALEACPQKRGVCTRVYTTTPKKPNSALRKVCRVRLTNGFEV
TSYIGGEGHNLQEHSVILIRGGRVKDLPGVRYHTVRGALDCSGVKDRKQARSKYGVKRPKA
>DisProt|DP00146|uniprot|P0A7T7|sp|RS18_ECOLI #1-75
MARYFRRRKFCRFTAEGVQEIDYKDIATLKNYITESGKIVPSRITGTRAKYQRQLARAIKRARYLS
LLPYTDRHQ
>DisProt|DP00147|uniprot|P0A7U3|sp|RS19_ECOLI #1-92
MPRSLKKGPFIDLHLLKKVEKAVESGDKKPLRTWSRRSTIFPNMIGLTIAVHNGRQHVPVFVTDE
MVGHKLGEFAPTRTYRGHAADKKAKKK
>DisProt|DP00148_C004|uniprot|P03347|sp|GAG_HV1B1 #1-55
MQRGNFRNQRKMVKCFNCGKEGHTARNCRAPRKKGCWKCGKEGHQMKDCTERQAN
>DisProt|DP00158|uniprot|P73124|sp|P73124_SYNY3 #1-65
MSTQQQARALMMRHHQFIKNRQQSMLSRAAAEIGVEAEKDFWTTVQGKPQSSFRTTYDRSNAS
LS
>DisProt|DP00164|uniprot|P05318|sp|RLA1_YEAST #1-106
MSTESALSYAALILADSEIEISSEKLLTLTNAANVPDENIWADIFAKALDGQNLKDLLVNFSAGAA
APAGVAGGVAGGEAGEAEAEKEEEEAKEESDDDMGFGLFD
>DisProt|DP00174|uniprot|P16949|unigene|Hs.209983|sp|STMN1_HUMAN #1-149
MASSDIQVKELEKRASGQAFELILSPRSKESVPEFPLSPPKKKDLSLEEIQKKLEAAEERRKSHEAE
VLKQLAEKREHEKEVLQKAIEENNNFSKMAEEKLTHKMEANKENREAQMAAKLERLREKDKHI
EEVRKNKESKDPADETEAD

>DisProt|DP00180_C003|uniprot|P19972|sp|TOXK_PICFA #1-77
GEATTIWGVGADEAIDKGTPSKNDLQNMSADLAKNGFKGHQGVACSTVKDGNKDVYMIKFSL
AGGSNDPGGSPCSDD
>DisProt|DP00185|uniprot|P93165|unigene|Gma.10|sp|P93165_SOYBN #1-105
MASRQNNKQELDERARQGETVVPGGTGGKSLEAQQHLAEGRSKGGQTRKEQLGTEGYQEMGR
KGGLSTVDKSGEERAQEEGIGIDESKFRTGNNKNQNQNEDQDK
>DisProt|DP00186|uniprot|Q95V77|sp|LEA1_APHAV #1-143
MSSQQNQNRQGEQQEQGYMEAAKEKVVNAWESTKETLSSTAQAAAEKTAEFRDSAGETIRDLT
GQAQEKGQEFKERAGEKAEETKQRAGEKMDETKQRAGEMRENAGQKMEEYKQQGKGKAEEL
RDTAAEKLHQAGEKVKGRD
>DisProt|DP00205|uniprot|Q82S91|sp|SMBP_NITEU #1-117
MKTTLIKVIAASVTALFLSMQVYASGHTAHVDEAVKHAEEAVAHGKEGHTDQLLEHAKESLTH
AKAASEAGGNTHVGHGIKHLEDAIKHGEEGHVGVATKHAQEAIEHLRASEHKSH
>DisProt|DP00216|uniprot|Q9FUM5|sp|Q9FUM5_BRANA #1-65
MADNKQSFQAGQAAGRAEEKGNVLMDKVKDAATAAGASAQTAGQKITEAAGGAVNLVKEKT
GMNK
>DisProt|DP00219|uniprot|O60927|unigene|Hs.82887|sp|PP1RB_HUMAN #1-126
MAEAGAGLSETVTETTVTVTTEPENRSLTIKLRKRKPEKKVEWTSDTVDNEHMGRRSSKCCCIY
EKPRAFGESSTESDEEEEEGCGHTHCVRGHRKGRRRATLGPTPTTPPQPPDPSQPPPGPMQH
>DisProt|DP00242|uniprot|P0AG63|sp|RS17_ECOLI #1-83
TDKIRTLQGRVVSDKMEKSIVVAIERFVKHPIYGKFIKRTTKLHVHDENNECGIGDVVEIRECRPL
SKTKSWTLVRVVEKAVL
>DisProt|DP00288|uniprot|Q06253|sp|PHD_BPP1 #1-73
MQSINFRTARGNLSEVLNNVEAGEEVEITRRGREPAVIVSKATFEAYKKAALDAEFASLFDTLDS
TNKELVNR
>DisProt|DP00347|uniprot|P04972|unigene|Bt.54|sp|CNRG_BOVIN #1-87
MNLEPPKAEIRSATRVMGGPVTPRKGPPKFKQRQTRQFKSKPPKKGVQGFGDDIPGMEGLGTDI
TVICPWEAFNHLELHELAQYGII
>DisProt|DP00357|uniprot|P62328|unigene|Hs.522584|sp|TYB4_HUMAN #1-44
MSDKPDMAEIEKFDKSKLKKTETQEKNPLPSKETIEQEKQAGES
>DisProt|DP00372|uniprot|Q9NR00|unigene|Hs.591849|sp|CH004_HUMAN #1-106
MKAKRSHQAIIMSTSLRVSPSIHGYHFDTASRKKAVGNIFENTDQESLERLFRNSGDKKAEERAKI
IFAIDQDVEEKTRALMALKKRTKDKLFQFLKLRKYSIKVH
>DisProt|DP00387|uniprot|P25814|sp|RNPA_BACSU #1-116
MKKRNRLKKNEDFQKVFKHGTSVANRQFVLYTLDQPENDELRVGLSVSKKIGNAVMRNRIKRL
IRQAFLEEKERLKEKDYIIIARKPASQLTYEETKKSLQHLFRKSSLYKKSSSK
>DisProt|DP00465|uniprot|Q57696|sp|Y246_METJA #1-99
MIEKLAEIRKKIDEIDNKILKLIAERNSLAKDVAEIKNQLGIPINDPEREKYIYDRIRKLCKEHNVDE
NIGIKIFQILIEHNKALQKQYLEETQNKNKK

>DisProt|DP00510|uniprot|O60356|unigene|Hs.513463|sp|NUPR1_HUMAN #1-82
MATFPPATSAPQQPPGPEDEDSSLDESDLYSLAHSYLGGGGRKGRTKREAAANTNRPSPGGHER
KLVTKLQNSERKKRGARR
>DisProt|DP00531|uniprot|Q08655|unigene|Les.17636|sp|ASR1_SOLLC #1-115
MEEEKHHHHHLFHHKDKAEEGPVDYEKEIKHHKHLEQIGKLGTVAAGAYALHEKHEAKKDPE
HAHKHKIEEEIAAAAAVGAGGFAFHEHHEKKDAKKEEKKKLRGDTTISSKLLF
>DisProt|DP00532|uniprot|Q8GT36|sp|Q8GT36_SPIOL #1-103
MSSLPFVFGAAASSRVVTAAAAKGTAETKQEKSFVDWLLGKITKEDQFYETDPILRGGDVKSSG
STSGKKGGTTSGKKGTVSIPSKKKNGNGGVFGGLFAKKD
>DisProt|DP00538|uniprot|A8CDV5|sp|A8CDV5_EBVG #1-118
MGSLEMVPMGAGPPSPGGDPDGDDGGNNSQYPSASGSSGNTPTPPNDEERESNEEPPPPYEDLD
WGNGDRHSDYQPLGNQDPSLYLGLQHDGNDGLPPPPYSPRDDSSQHIYEEAGRG
>DisProt|DP00544|uniprot|B0FRH7|sp|LLPH_APLKU #1-120
MAKSIRSKHRRQMRNVKREHFAKKDLDRLKRLASKAQELDLDNVVTMKSAEEIKNKPSTSASD
ADKGMEVDNTKKVFKKKTQQNEDGHYPQWMNQRAVKKQKVKVAKLKTKKKIGKKIKW
>DisProt|DP00550|uniprot|P02628|sp|PRVA_ESOLU #1-108
AKDLLKADDIKKALDAVKAEGSFNHKKFFALVGLKAMSANDVKKVFKAIDADASGFIEEEELKF
VLKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHEA
>DisProt|DP00555|uniprot|Q16143|unigene|Hs.90297|sp|SYUB_HUMAN #1-134
MDVFMKGLSMAKEGVVAAAEKTKQGVTEAAEKTKEGVLYVGSKTREGVVQGVASVAEKTKE
QASHLGGAVFSGAGNIAAATGLVKREEFPTDLKPEEVAQEAAEEPLIEPLMEPEGESYEDPPQEE
YQEYEPEA
>DisProt|DP00586|uniprot|P01094|sp|IPA3_YEAST #1-68
MNTDQQKVSEIFQSSKEKLQGDAKVVSDAFKKMASQDKDGKTTDADESEKHNYQEQYNKLKG
AGHKKE
>DisProt|DP00592|uniprot|P48539|unigene|Hs.80296|sp|PCP4_HUMAN #1-62
MSERQGAGATNGKDKTSGENDGQKKVQEEFDIDMDAPETERAAVAIQSQFRKFQKKKAGSQS
>DisProt|DP00626|uniprot|P0AG11|sp|UMUD_ECOLI #1-139
MLFIKPADLREIVTFPLFSDLVQCGFPSPAADYVEQRIDLNQLLIQHPSATYFVKASGDSMIDGGIS
DGDLLIVDSAITASHGDIVIAAVDGEFTVKKLQLRPTVQLIPMNSAYSPITISSEDTLDVFGVVIHV
VKAMR
>DisProt|DP00630|uniprot|O76070|unigene|Hs.349470|sp|SYUG_HUMAN #1-127
MDVFKKGFSIAKEGVVGAVEKTKQGVTEAAEKTKEGVMYVGAKTKENVVQSVTSVAEKTKEQ
ANAVSEAVVSSVNTVATKTVEEAENIAVTSGVVRKEDLRPSAPQQEGEASKEKEEVAEEAQSGG
D
>DisProt|DP00650|uniprot|Q1PAB4|sp|Q1PAB4_9HIV1 #1-101
MEPVDPRLEPWKHPGSQPRTACTNCYCKKCCFHCQVCFIRKALGISYGRKKRRQRRRAPQDSET
HQVSPPKQPASQPRGDPTGPKESKKKVERETETHPVN

>DisProt|DP00665|uniprot|Q9XES8|unigene|Gma.168|sp|Q9XES8_SOYBN #1-89

MAKSKEDITYATSQARLSEDEAVRVAYEHGSPLEGGKIADSQPVDLFSSAHNMPKSGQTTMDSN

TSDQSQMQRDTQEGGSKEFTTGAPG

>DisProt|DP00675_C002|uniprot|P19711|sp|POLG_BVDVN #1-102

SDTKEEGATKKKTQKPDRLERGKMKIVPKESEKDSKTKPPDATIVVEGVKYQVRKKGKTKSKN

TQDGLYHNKNKPQESRKKLEKALLAWAIIAIVLFQVTMG

ANNEX III

ArchCandy: arc- and position-specific rules

Numbering:

Numbers without '[]' show the residue number with respect to the sequence. Numbers with '[]' are internal residues.

Abbreviations:

Ar: Aromatic residues: W, F, Y

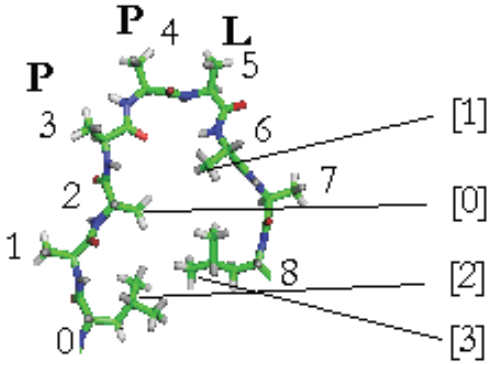
Me: Medium sized residues: L, M, I, R, H, K, D, E, T, C, P , V, N , Q

Sm: Small sized residues: S, A, G

Residue Conformations: p - polyproline; a – alpha-helical, b – beta-structural; l – left-handed alpha-helical; g – 3_{10}-helical; e- glycine-specific.

1. Arc Steric Tension Score:

3-residue arch (ppl-conformation)



a. If in [0] and [1] are L, I, Y, F or W = NO (score=0)

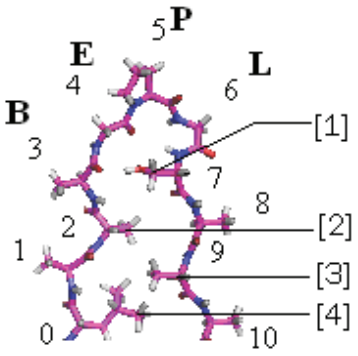
b. If in [1] is W and in [0] is not G or A = NO (score=0)

c. If in one position is G, A or S and in the other Any residue, score = 1.0

d. If in both positions are no G, A or S, nor L, I, F, Y, W, score = 0.8

e. If in one position no G, A, S but in the other L, I, F, Y, W, score = 0.6

4-residue-arch (bepl-conformation



1. **Glycine:** must be present at pos 4
2. **Prolines** are allowed at positions 3 and 5
3. **Steric Constraints**

a) 2 Ar in positions [1] and [2]; score=0

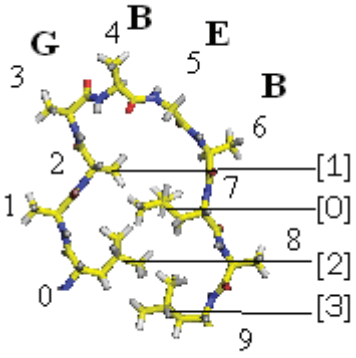
b) 1 Ar:

(i) If [1] is W; score=0.0

(ii) If [1] is Y; score=0.8

(iii) If [1] is F; score=0.9

4-residue-arch (gbeb-conformation)



1.**Prolines:** Only allowed at positions 3, 4, and 6 relative to the sequence.

2.**Glycine:** Position 5 has to be G

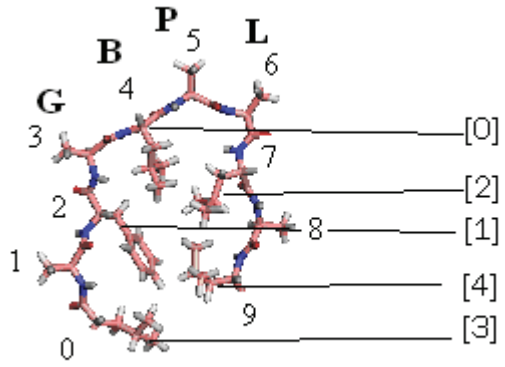
3.**Steric Constraints:** (For positions [0] and [1] in diagram)

a) 2 Ar; score=0

b) 1Ar, 1Me: If [1] is Ar then score=0.

c) Everything else; score=1

4-residue arch (gbpl-conformation)



---

1. **Prolines:** Not allowed at any position.
2. **Steric Constraints:** (For positions [0] [1] [2] in diagram)

a) 3Ar=0

b) 2Ar, 1Me=0

c) 2Ar,1Sm:

(i) If '1W+1F' occur at any position, score=0.7

(ii) If '1W+1Y' occur at any position, score=0.7

(iii)If '2W', score=0

d) 1Ar, 2Me:

(i) If pos [0] is 'W' score=0,

(ii) Else score=0.8

e) 1Ar,1Me,1Sm:

(i) If 1W occurs at any position, score=0.8

(ii) If 1Y occurs at any position, score=0.9

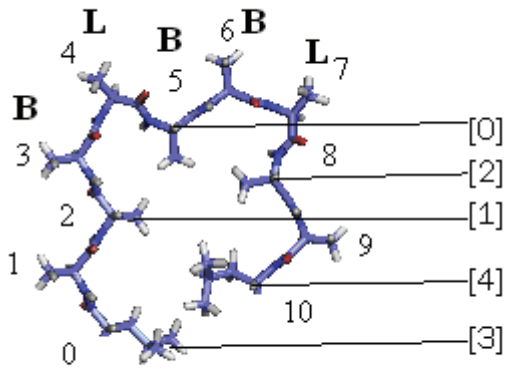
(iii) Else score=1

f) 3Me:

(i) If pos [0] is 'L', score=0

(ii) Else score=0.9

5-residue arch



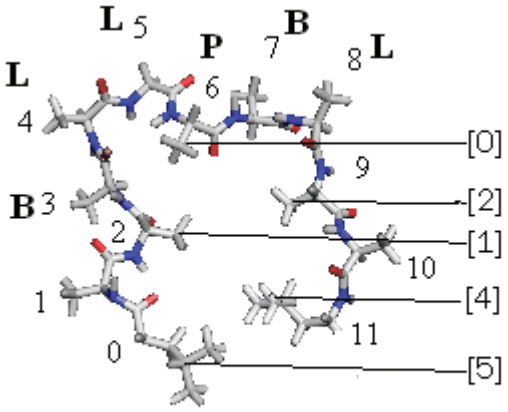
---

1. **Prolines:** Prolines are not allowed on beta-strands but they are allowed in arc region. However, some positions called "intermediate 1 and 2" have penalties. The "intermediate 2" positions are 4 and 7. Prolines in these positions reduce the score by 0.65^n where n is the number of prolines in the "intermediate 2" positions. "Intermediate 1" are positions: 3, 5, and 6. If a proline occurs in any of these positions the score is reduced by 0.9^n.
2. **Steric Constraints:** (For pos [0], [1] and [2] in diagram)

a. Ar are >=2, score=0

b. Everything else, score =1

6-residue arch type 1

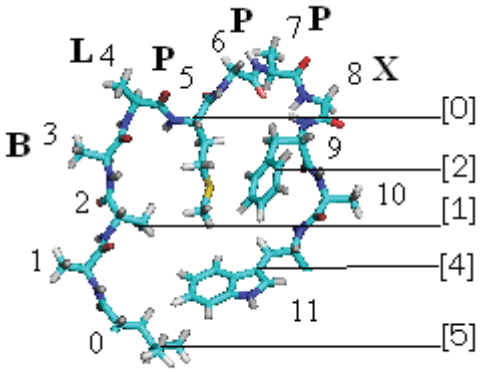


---

1. **Prolines:** Prolines are not allowed on beta-strands but they are allowed in arc region. However, some positions called "intermediate 1 and 2" have penalties. The "intermediate 2" positions are 4, 5 and 8. Prolines in these positions reduce the score by $(0.65)^n$ where n is the number of prolines in the "intermediate 2" positions. "Intermediate 1" are positions: 3, 6, and 7. If a proline occurs in any of these positions the score is reduced by 0.9^n.
2. Steric Tension with Gly at position 5: 3Ar, Score=0; everything else is allowed.

a. **Steric Tension:**
b. 3Ar, score=0
c. 2Ar, if both pos [1] and [0] are Ar, score=0
d. 1Ar, 2Me, if pos [0] is "W", score=0
e. 1Ar, 1Me ,1s, if pos [0] is "W", score=0
f. Everything else, score=1

6-residue arch type 2



---

1. **Prolines:** Prolines are not allowed on beta-strands but they are allowed in arc region. However, some positions called “intermediate 1 and 2” have penalties. The “intermediate 2” positions are 4, 6 and 8. Prolines in these positions reduce the score by $(0.65)^n$ where n is the number of prolines in the “intermediate 2” positions. “Intermediate 1” are positions: 3, 5, and 7. If a proline occurs in any of these positions the score is reduced by (0.9).
2. **Special:** Position 6 must be ‘A’ or ‘G’
3. **Steric Tension** in all other cases:

a. 3Ar, Score=0

b. 2Ar if pos [0], and [2] are Ar and [1] is not V, H, C, S, A, or G; score=0

c. Everything smaller is allowed.

2. Glycine in Arc Score

Number of glycines in the arc	Glycine in Arc Score
0	0.8
>0	1.00

3. Arc Length Score

Weight assigned to each type of arc.

6 residue arcs: 0.85

5 residue arcs: 0.95

4 residue arcs: 1.0

3 residue arcs: 1.0

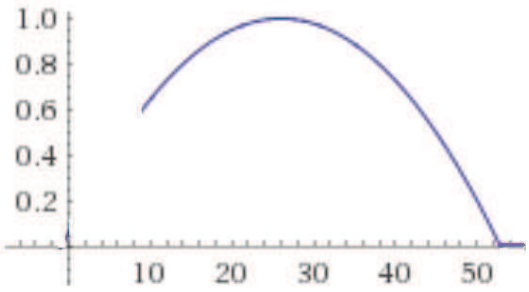
4. Beta-Strand Length Score

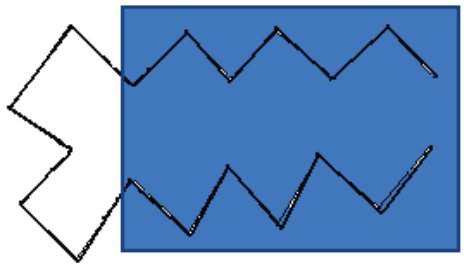
Score = 0.61 when L=10; otherwise

Score = 1- $[0.0003462*(2L-lmin-lmax)^2]$

Where **L** is total length of both the beta-strands (total arch length – arc length) and lmin=7, lmax=45.

Max score =1.0 is at L=26.

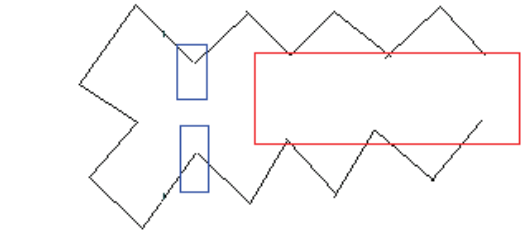




Arc Region B-Strand region

5. Internal AA Composition Score

Diagram to explain key hydrophobic positions:



Internal AA Composition Score = 1 – {[badIR+(alaIR*0.2) +(thrIR*0.8)+(internalK*0.3)+(internalR*0.5)]/totalIR

Where:

- badIR: Number of "unfavourable" residues in key internal hydrophobic positions that decrease the stability of the fibril. Amino acids considered "unfavourable" are: Ser,His,Cis,Gly
- alaIR: Number of Ala in key internal hydrophobic positions.
- thrIR: Number of Thr in key internal hydrophobic positions.
- internalK: Number of Lys involved in salt-bridges in the whole hydrophobic region, not just the key region.
- internalR: Number of Arg involved in the salt-bridges in the whole hydrophobic region, not just the key region.
- totalIR: Total number of internal residues in key hydrophobic positions

6. Total-Net-Charge Score

Total-Net-Charge Score = e^x

Where:

- x = -4*(netCharge)^2
- netCharge = |kr-de|/sequenceLength
- (absolute difference between positive and negative charge divided by the total length of the sequence)
- kr: total number of 'K' or 'R' residues in the candidate sequence
- de: total number of 'D' or 'E' residues in the candidate sequence

7. Proportion of Charged Residue Score

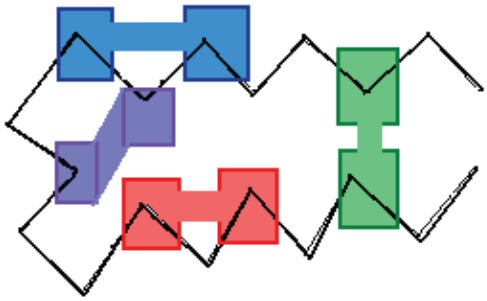
Score= $e^{-1.25*PropChargedRes}$

PropChargedRes = (total charged residues in candidate that are not involved in salt-bridges)/(candidate sequence length).

8. Internal Salt-Bridge Score

SideIonicBonds is the total number of internal salt bridges on the same beta-strand of the candidate.

If **SideIonicBonds**>=2, **Internal Salt-Bridge Score=$(0.65)^2$**
If **SideIonicBonds**=1, **Internal Salt-Bridge Score=0.65**
If **SideIonicBonds**=0, **Internal Salt-Bridge Score=1.0**



External Salt Bridges are formed between two neighbouring external residues.

Internal "cross" Salt bridges are formed between two internal residues on different B-strands.

Internal "Side" Salt Bridges are formed between two internal residues on the same B-strand.

Arc Salt bridges are formed between an internal residue in the arc and either one of the two first internal residues on each B-strand.

ANNEX IV

ArchCandy Extended Dataset

Postive set

> Human B2-microglobulin Mutant fragment (PDB code: 2E8D)

SNFLNCYVSGFHPSDIEVDLLK

> Human CA150 (PDB code: 2NNT)

MGATAVSEWTEYKTADGKTFYYNNRTLESTW

> HET-s Prion from Saccharomyces cerevisiae (218-289)

KIDAIVGRNSAKDIRTEERARVQLGNVVTAAALHGGIRISDQTTNSVETVVGKGESRVLIGNEYG
GKGFWDN

> Human calcitonin

CGNLSTCMLGTTTQDFNLFHTFPQTAIGVGAP

> Human Semen-derived Enhancer of Viral Infection (SEVI) Fibril Forming peptide of Prostatic Acid Phosphatase Peptide (248-286)

YGIHKQKEKSRLQGGVLVNEILNHMKRATQIPSYKKLIMY

> Sup35 from Saccharomyces cerevisiae (1-114)

MSDSNQGNNQQNYQQYSQNGNQQQGNNRYQGYQAYNAQAQPAGGYYQNYQGYSGYQQGG
YQQYNPDAGYQQQYNPQGGYQQYNPQGGYQQQFNPQGGRGNYKNFNYNNNLQGYQ

> Ure2P from Saccharomyces cerevisiae (1-94)

MMNNNGNQVSNLSNALRQVNIGSRNSNTTTDQSNINFEFSTGVNNNNNNNSSSNNNNVQNNNS
GRNGSQNNDNENNIKNTLEQHRQQQQAFSDM

> Rnq1p from Saccharomyces cerevisiae (153-405)

QGQGQGQGQGQGQGQGQGQGSFTALASLASSFMNSNNNNQQGQNQSSGGSSFGALASMASSF
MHSNNNQNSNNSQQGYNQSYQNGNQNSQGYNNQQYQGGNGGYQQQQGQSGGAFSSLASMA
QSYLGGGQTQSNQQQYNQQGQNNQQQYQQQGQNYQHQQQGQQQQQGHSSSFSALASMASSY
LGNNSNSNSSYGGQQQANEYGRPQQNGQQQSNEYGRPQYGGNQNSNGQHESFNFSGNFSQQN
NNGNQNRY

> Human Ataxin Diseases (including huntingtin)

QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ

> Streptomyces coelicolor Chaplin F

DSGAQAAAAHSPGVLSGNVVQVPVHIPVNVCGNTIDVIGLLNPAFGNECEND

> Streptomyces coelicolor Chaplin H

DSGAQGAAVHSPGVLSGNVVQVPVHVPVNVCGNTISVIGLLNPAFGNVCINK

> Streptomyces coelicolor Chaplin G

DAGAAGAAVGSPGVLSGNVVQVPVHVPVNlCGNTIDVIGLLNPAFGNACENGDDDKSGGYGG

> Streptomyces coelicolor Chaplin D

DAGAEGAAVGSPGVLSGNVIQVPVHVPVNVCGNSINVVGLLNPAFGNKCEND

> Streptomyces coelicolor Chaplin E

TDGGAHAHGKAVGSPGVASGNLVQAPIHIPVNAVGNSISVNVIGVLNPAFGNLGVNH

> Microcin E492 from Klebsiella pneumoniae (16-99)

GETDPNTQLLNDLGNNMAWGAALGAPGGLGSAALGAAGGALQTVGQGLIDHGPVNVPIPVLIG
PSWNGSGSGYNSATSSSGSGS

> Prion Formation Protein 1 from from Saccharomyces cerevisiae (1-100)

MPPKKFKDLNSFLDDQPKDPNLVASPFGGYFKNPAADAGSNNASKKSSYQQQRNWKQGGNYQ
QGGYQSYNSNYNNYNNYNNYNNYNNYNNYNKYNGQGYQ

> Human RIP1 (519-560)

SSLPPTDESIKYTIYNSTGIQIGAYNYMEIGGTSSSLLDST

> Human RIP3 (439-479)

PEPNPVTGRPLVNIYNCSGVQVGDNNYLTMQQTTALPTWGL

> Human TDP(TAR DNA-binding Protein) (281-332)

GFGNSRGGGAGLGNNQGSNMGGGMNFGAFSINPAMMAAAQAALQS

> Human Prp (23-230)

KKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGGWGQPHGGGWG
QPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGGLGGYMLGSAMSRPIIHFGS
DYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKM
MERVVEQMCITQYERESQAYYQRGS

> Human amyloid-beta40 (AB40)

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV

> Human amyloid-beta42 (AB42)

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA

>Human amyloid-beta43 (AB43)

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIAT

>Human AB40||A21G Flemish mutation

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVGGVV

>Human AB40||E22G Artic mutation

DAEFRHDSGYEVHHQKLVFFAGDVGSNKGAIIGLMVGGVV

>Human AB40||E22Q Dutch mutation

DAEFRHDSGYEVHHQKLVFFAQDVGSNKGAIIGLMVGGVV

>Human AB40||E22K Italian mutation

DAEFRHDSGYEVHHQKLVFFAKDVGSNKGAIIGLMVGGVV

>Human AB40||D23N Iowa mutation

DAEFRHDSGYEVHHQKLVFFAENVGSNKGAIIGLMVGGVV

>Human AB40||T43I Austrian mutation

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIAI

>Human AB40||E22del Japanese mutation

DAEFRHDSGYEVHHQKLVFFADVGSNKGAIIGLMVGGVV

>Human AB40||A21G Flemish mutation

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVGGVV

>Human AB40||H14P

DAEFRHDSGYEVPHQKLVFFGEDVGSNKGAIIGLMVGGVV

>Human AB40||E22P

DAEFRHDSGYEVHHQKLVFFGPDVGSNKGAIIGLMVGGVV

>Human AB40||D23P

DAEFRHDSGYEVHHQKLVFFGEPVGSNKGAIIGLMVGGVV

>Human AB40||G29P

DAEFRHDSGYEVHHQKLVFFGEDVGSNKPAIIGLMVGGVV

>Human AB40||A30P

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGPIIGLMVGGVV

>Human AB40||G37P

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVPGVV

>Human AB40||G38P

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVGPVV

>Human AB40||V39P

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVGGPV

> Human amylin

KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY

>Amylin||hIAPP 3xL

KCNTATCATQRLANLLVHSSNNLGAILSSTNVGSNTL

>Amylin||S28G

KCNTATCATQRLANFLVHSSNNFGAILGSTNVGSNTY

>Amylin||S28K

KCNTATCATQRLANFLVHSSNNFGAILKSTNVGSNTY

>Amylin||S20G

KCNTATCATQRLANFLVHSGNNFGAILSSTNVGSNTY

> Human alpha-synuclein

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVSKTKEGVVHGVATVAEKTKEQ
VTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEA
YEMPSEEGYQDYEPEA

>a-synuclein E46K

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKKGVLYVGSKTKEGVVHGVATVAEKTKE
QVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA

>a-synuclein A53T

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKKGVLYVGSKTKEGVVHGVTTVAEKTKE
QVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA

>a-synuclein A30P

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAPGKTKKGVLYVGSKTKEGVVHGVTTVAEKTKE
QVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA

>serum amyloid A-2 protein isoform SAA2.2, Mus musculus (20-122)

GFFSFVHEAFLGAGDMWRAYTDMKEAGWKDGDKYFHARGNYDAAQRGPGGVWAAEKISDG
REAFQEFFGRGHEDTMADQEANRHGRSGKDPNYYRPPGLPDKY

>sp|P28307|21-151 CsgA E. coli K12

GVVPQYGGGGNHGGGGNNSGPNSELNIYQYGGGNSALALQTDARNSDLTITQHGGGNGADVG
QGSDDSSIDLTQRGFGNSATLDQWNGKNSEMTVKQFGGGNGAAVDQTASNSSVNVTQVGFGN
NATAHQY

>sp|P0ABK7|22-151 CsgB E. coli K12

AGYDLANSEYNFAVNELSKSSFNQAAIIGQAGTNNSAQLRQGGSKLLAVVAQEGSSNRAKIDQT
GDYNLAYIDQAGSANDASISQGAYGNTAMIIQKGSGNKANITQYGTQKTAIVVQRQSQMAIRVT
QR

>sp|P40967|25-467 Pmel17 Malpha domain

KVPRNQDWLGVSRQLRTKAWNRQLYPEWTEAQRLDCWRGGQVSLKVSNDGPTLIGANASFSIA
LNFPGSQKVLPDGQVIWVNNTIINGSQVWGGQPVYPQETDDACIFPDGGPCPSGSWSQKRSFVY
VWKTWGQYWQVLGGPVSGLSIGTGRAMLGTHTMEVTVYHRRGSRSYVPLAHSSSAFTITDQVP
FSVSVSQLRALDGGNKHFLRNQPLTFALQLHDPSGYLAEADLSYTWDFGDSSGTLISRALVVTHT
YLEPGPVTAQVVLQAAIPLTSCGSSPVPGTTDGHRPTAEAPNTTAGQVPTTEVVGTTPGQAPTAE
PSGTTSVQVPTTEVISTAPVQMPTAESTGMTPEKVPVSEVMGTTLAEMSTPEATGMTPAEVSIVV
LSGTTAAQVTTTEWVETTARELPIPEPEGPDASSIMSTESITGSLGPLLDGTATLRLV

Negative Dataset

>Human AB40||F19P -:

DAEFRHDSGYEVHHQKLVPFGEDVGSNKGAIIGLMVGGVV

>Human AB40||F20P -:

DAEFRHDSGYEVHHQKLVFPGEDVGSNKGAIIGLMVGGVV

>Rat amylin (no fibrils in vivo)

KCNTATCATQRLANFLVRSSNNLGPVLPPTNVGSNTY

>Hamster amylin (no fibrils in vivo)

KCNTATCATQRLANFLVHSNNNFGPVLSPTNVGSNTY

>Degu amylin (no fibrils in vivo)

KCNTATCATQRLTNFLVRSSHNLGAALPPTKVGSNTY

>Human Amylin||I26D -

KCNTATCATQRLANFLVHSSNNFGADLSSTNVGSNTY

>Human Amylin||A13E -

KCNTATCATQRLENFLVHSSNNFGAILSSTNVGSNTY

>Human Amylin||L16Q -

KCNTATCATQRLANFQVHSSNNFGAILSSTNVGSNTY

>Human Amylin||hIAPP 8-37 3xP -

ATQRLANFLPHPSNNFGAILSSPNVGSNTY

>Human Amylin||N22P -

KCNTATCATQRLANFLVHSSNPFGAILSSTNVGSNTY

>Human Amylin||G24P -

KCNTATCATQRLANFLVHSSNNFPAILSSTNVGSNTY

>Human Amylin||I26P -

KCNTATCATQRLANFLVHSSNNFGAPLSSTNVGSNTY

>Human Amylin||L27P -

KCNTATCATQRLANFLVHSSNNFGAIPSSTNVGSNTY

>S28P -

KCNTATCATQRLANFLVHSSNNFGAILPSTNVGSNTY

>a-synuclein A76E -:

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAPGKTKKGVLYVGSKTKEGVVHGVTTVAEKTK
EQVTNVGGAVVTGVTEVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA

>a-synuclein A76R -:

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAPGKTKKGVLYVGSKTKEGVVHGVTTVAEKTK
EQVTNVGGAVVTGVTRVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA

>DisProt|DP00001|uniprot|Q9HFQ6|sp|RLA3_CANAL #1-108

MSTEASVSYAALILADAEQEITSEKLLAITKAAGANVDQVWADVFAKAVEGKNLKELLFSFAA
AAPASGAAAGSASGAAAGGEAAAEEAAEEEAAEESDDDMGFGLFD

>DisProt|DP00002|uniprot|P02400|sp|RLA4_YEAST #1-110

MKYLAAYLLLVQGGNAAPSAADIKAVVESVGAEVDEARINELLSSLEGKGSLEEIIAEGQKKFA
TVPTGGASSAAAGAAGAAAGGDAAEEEKEEEAKEESDDDMGFGLFD

>DisProt|DP00004_C002|uniprot|P49913|unigene|Hs.51120|sp|CAMP_HUMAN #1-37

LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES

>DisProt|DP00005|uniprot|P03045|sp|REGN_LAMBD #1-107

MDAQTRRRERRAEKQAQWKAANPLLVGVSAKPVNLPILSLNRKPKSRVESALNPIDLTVLAEY
HKQIESNLQRIERKNQRTWYSKPGERGITCSGRQKIKGKSIPLI

>DisProt|DP00006|uniprot|P00004|unigene|Eca.1571|sp|CYC_HORSE #1-104

GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEE
TLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE

>DisProt|DP00022|uniprot|P17639|sp|EMB1_DAUCA #1-92

MASQQEKKELDARARQGETVVPGGTGGKSLEAQQHLAEGRSKGGQTRKEQLGGEGYHEMGR
KGGLSNNDMSGGERAEQEGIDIDESKFRTKK

>DisProt|DP00024|uniprot|P03129|sp|VE7_HPV16 #1-98

MHGDTPTLHEYMLDLQPETTDLYCYEQLSDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCKCDST
LRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP
>DisProt|DP00027|uniprot|P26477|sp|FLGM_SALTY #1-97
MSIDRTSPLKPVSTVQTRETSDTPVQKTRQEKTSAATSASVTLSDAQAKLMQPGVSDINMERVE
ALKTAIRNGELKMDTGKIADSLIREAQSYLQSK
>DisProt|DP00028|uniprot|Q13541|unigene|Hs.411641|sp|4EBP1_HUMAN #1-118
MSGGSSCSQTPSRAIPATRRVVLGDGVQLPPGDYSTTPGGTLFSTTPGGTRIIYDRKFLMECRNSP
VTKTPPRDLPTIPGVTSPSSDEPPMEASQSHLRNSPEDKRAGGEESQFEMDI
>DisProt|DP00039|uniprot|P05204|unigene|Hs.181163|sp|HMGN2_HUMAN #1-89
PKRKAEGDAKGDKAKVKDEPQRRSARLSAKPAPPKPEPKPKKAPAKKGEKVPKGKKGKADAG
KEGNNPAENGDAKTDQAQKAEGAGDAK
>DisProt|DP00040|uniprot|P17096|unigene|Hs.518805|sp|HMGA1_HUMAN #1-107
MSESSSKSSQPLASKQEKDGTEKRGRGRPRKQPPVSPGTALVGSQKEPSEVPTPKRPRGRPKGSK
NKGAAKTRKTTTTPGRKPRGRPKKLEKEEEEGISQESSEEEQ
>DisProt|DP00057|uniprot|P15340|sp|HSP1_CHICK #1-62
MARYRRSRTRSRSPRSRRRRRRSGRRRSPRRRRRYGSARRSRRSVGGRRRRYGSRRRRRRRY
>DisProt|DP00058|uniprot|P06302|unigene|Rn.817|sp|PTMA_RAT #1-112
MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAQNEENGEQEADNEVDEEEEEGGEE
EEEEEEGDGEEEDGDEDEEAEAPTGKRVAEDDEDDDVETKKQKKTDEDD
>DisProt|DP00070|uniprot|P37840-1|unigene|Hs.21374|sp|SYUA_HUMAN #1-140
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTK
EQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNE
AYEMPSEEGYQDYEPEA
>DisProt|DP00116|uniprot|P81455|sp|OSTCN_CANFA #1-49
YLDSGLGAPVPYPDPLEPKREVCELNPNCDELADHIGFQEAYQRFYGPV
>DisProt|DP00140|uniprot|P0A7L8|sp|RL27_ECOLI #1-85
MAHKKAGGSTRNGRDSEAKRLGVKRFGGESVLAGSIIVRQRGTKFHAGANVGCGRDHTLFAK
ADGKVKFEVKGPKNRKFISIEAE
>DisProt|DP00143|uniprot|P0A7N9|sp|RL33_ECOLI #1-55
MAKGIREKIKLVSSAGTGHFYTTTKNKRTKPEKLELKKFDPVVRQHVIYKEAKIK
>DisProt|DP00145|uniprot|P0A7S3|sp|RS12_ECOLI #1-124
MATVNQLVRKPRARKVAKSNVPALEACPQKRGVCTRVYTTTPKKPNSALRKVCRVRLTNGFE
VTSYIGGEGHNLQEHSVILIRGGRVKDLPGVRYHTVRGALDCSGVKDRKQARSKYGVKRPKA
>DisProt|DP00146|uniprot|P0A7T7|sp|RS18_ECOLI #1-75
MARYFRRRKFCRFTAEGVQEIDYKDIATLKNYITESGKIVPSRITGTRAKYQRQLARAIKRARYL
SLLPYTDRHQ
>DisProt|DP00147|uniprot|P0A7U3|sp|RS19_ECOLI #1-92
MPRSLKKGPFIDLHLLKKVEKAVESGDKKPLRTWSRRSTIFPNMIGLTIAVHNGRQHVPVFVTD
EMVGHKLGEFAPTRTYRGHAADKKAKKK
>DisProt|DP00148_C004|uniprot|P03347|sp|GAG_HV1B1 #1-55

MQRGNFRNQRKMVKCFNCGKEGHTARNCRAPRKKGCWKCGKEGHQMKDCTERQAN

>DisProt|DP00158|uniprot|P73124|sp|P73124_SYNY3 #1-65

MSTQQQARALMMRHHQFIKNRQQSMLSRAAAEIGVEAEKDFWTTVQGKPQSSFRTTYDRSNA
SLS

>DisProt|DP00164|uniprot|P05318|sp|RLA1_YEAST #1-106

MSTESALSYAALILADSEIEISSEKLLTLTNAANVPDENIWADIFAKALDGQNLKDLLVNFSAGA
AAPAGVAGGVAGGEAGEAEAEKEEEEAKEESDDDMGFGLFD

>DisProt|DP00174|uniprot|P16949|unigene|Hs.209983|sp|STMN1_HUMAN #1-149

MASSDIQVKELEKRASGQAFELILSPRSKESVPEFPLSPPKKKDLSLEEIQKKLEAAEERRKSHEA
EVLKQLAEKREHEKEVLQKAIEENNNFSKMAEEKLTHKMEANKENREAQMAAKLERLREKDK
HIEEVRKNKESKDPADETEAD

>DisProt|DP00180_C003|uniprot|P19972|sp|TOXK_PICFA #1-77

GEATTIWGVGADEAIDKGTPSKNDLQNMSADLAKNGFKGHQGVACSTVKDGNKDVYMIKFSL
AGGSNDPGGSPCSDD

>DisProt|DP00185|uniprot|P93165|unigene|Gma.10|sp|P93165_SOYBN #1-105

MASRQNNKQELDERARQGETVVPGGTGGKSLEAQQHLAEGRSKGGQTRKEQLGTEGYQEMG
RKGGLSTVDKSGEERAQEEGIGIDESKFRTGNNKNQNQNEDQDK

>DisProt|DP00186|uniprot|Q95V77|sp|LEA1_APHAV #1-143

MSSQQNQNRQGEQQEQGYMEAAKEKVVNAWESTKETLSSTAQAAAEKTAEFRDSAGETIRDL
TGQAQEKGQEFKERAGEKAEETKQRAGEKMDETKQRAGEMRENAGQKMEEYKQQGKGKAEE
LRDTAAEKLHQAGEKVKGRD

>DisProt|DP00205|uniprot|Q82S91|sp|SMBP_NITEU #1-117

MKTTLIKVIAASVTALFLSMQVYASGHTAHVDEAVKHAEEAVAHGKEGHTDQLLEHAKESLT
HAKAASEAGGNTHVGHGIKHLEDAIKHGEEGHVGVATKHAQEAIEHLRASEHKSH

>DisProt|DP00216|uniprot|Q9FUM5|sp|Q9FUM5_BRANA #1-65

MADNKQSFQAGQAAGRAEEKGNVLMDKVKDAATAAGASAQTAGQKITEAAGGAVNLVKEK
TGMNK

>DisProt|DP00219|uniprot|O60927|unigene|Hs.82887|sp|PP1RB_HUMAN #1-126

MAEAGAGLSETVTETTVTVTTEPENRSLTIKLRKRKPEKKVEWTSDTVDNEHMGRRSSKCCCIY
EKPRAFGESSTESDEEEEEGCGHTHCVRGHRKGRRRATLGPTPTTPPQPPDPSQPPPGPMQH

>DisProt|DP00242|uniprot|P0AG63|sp|RS17_ECOLI #1-83

TDKIRTLQGRVVSDKMEKSIVVAIERFVKHPIYGKFIKRTTKLHVHDENNECGIGDVVEIRECRP
LSKTKSWTLVRVVEKAVL

>DisProt|DP00288|uniprot|Q06253|sp|PHD_BPP1 #1-73

MQSINFRTARGNLSEVLNNVEAGEEVEITRRGREPAVIVSKATFEAYKKAALDAEFASLFDTLDS
TNKELVNR

>DisProt|DP00347|uniprot|P04972|unigene|Bt.54|sp|CNRG_BOVIN #1-87

MNLEPPKAEIRSATRVMGGPVTPRKGPPKFKQRQTRQFKSKPPKKGVQGFGDDIPGMEGLGTDI
TVICPWEAFNHLELHELAQYGII

>DisProt|DP00357|uniprot|P62328|unigene|Hs.522584|sp|TYB4_HUMAN #1-44

MSDKPDMAEIEKFDKSKLKKTETQEKNPLPSKETIEQEKQAGES

>DisProt|DP00372|uniprot|Q9NR00|unigene|Hs.591849|sp|CH004_HUMAN #1-106

MKAKRSHQAIIMSTSLRVSPSIHGYHFDTASRKKAVGNIFENTDQESLERLFRNSGDKKAEERA
KIIFAIDQDVEEKTRALMALKKRTKDKLFQFLKLRKYSIKVH

>DisProt|DP00387|uniprot|P25814|sp|RNPA_BACSU #1-116

MKKRNRLKKNEDFQKVFKHGTSVANRQFVLYTLDQPENDELRVGLSVSKKIGNAVMRNRIKR
LIRQAFLEEKERLKEKDYIIIARKPASQLTYEETKKSLQHLFRKSSLYKKSSSK

>DisProt|DP00465|uniprot|Q57696|sp|Y246_METJA #1-99

MIEKLAEIRKKIDEIDNKILKLIAERNSLAKDVAEIKNQLGIPINDPEREKYIYDRIRKLCKEHNVD
ENIGIKIFQILIEHNKALQKQYLEETQNKNKK

>DisProt|DP00510|uniprot|O60356|unigene|Hs.513463|sp|NUPR1_HUMAN #1-82

MATFPPATSAPQQPPGPEDEDSSLDESDLYSLAHSYLGGGGRKGRTKREAAANTNRPSPGGHER
KLVTKLQNSERKKRGARR

>DisProt|DP00531|uniprot|Q08655|unigene|Les.17636|sp|ASR1_SOLLC #1-115

MEEEKHHHHHLFHHKDKAEEGPVDYEKEIKHHKHLEQIGKLGTVAAGAYALHEKHEAKKDPE
HAHKHKIEEEIAAAAAVGAGGFAFHEHHEKKDAKKEEKKKLRGDTTISSKLLF

>DisProt|DP00532|uniprot|Q8GT36|sp|Q8GT36_SPIOL #1-103

MSSLPFVFGAAASSRVVTAAAAKGTAETKQEKSFVDWLLGKITKEDQFYETDPILRGGDVKSSG
STSGKKGGTTSGKKGTVSIPSKKKNGNGGVFGGLFAKKD

>DisProt|DP00538|uniprot|A8CDV5|sp|A8CDV5_EBVG #1-118

MGSLEMVPMGAGPPSPGGDPDGDDGGNNSQYPSASGSSGNTPTPPNDEERESNEEPPPPYEDLD
WGNGDRHSDYQPLGNQDPSLYLGLQHDGNDGLPPPPYSPRDDSSQHIYEEAGRG

>DisProt|DP00544|uniprot|B0FRH7|sp|LLPH_APLKU #1-120

MAKSIRSKHRRQMRNVKREHFAKKDLDRLKRLASKAQELDLDNVVTMKSAEEIKNKPSTSASD
ADKGMEVDNTKKVFKKKTQQNEDGHYPQWMNQRAVKKQKVKVAKLKTKKKIGKKIKW

>DisProt|DP00550|uniprot|P02628|sp|PRVA_ESOLU #1-108

AKDLLKADDIKKALDAVKAEGSFNHKKFFALVGLKAMSANDVKKVFKAIDADASGFIEEEELK
FVLKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHEA

>DisProt|DP00555|uniprot|Q16143|unigene|Hs.90297|sp|SYUB_HUMAN #1-134

MDVFMKGLSMAKEGVVAAAEKTKQGVTEAAEKTKEGVLYVGSKTREGVVQGVASVAEKTK
EQASHLGGAVFSGAGNIAAATGLVKREEFPTDLKPEEVAQEAAEEPLIEPLMEPEGESYEDPPQE
EYQEYEPEA

>DisProt|DP00586|uniprot|P01094|sp|IPA3_YEAST #1-68

MNTDQQKVSEIFQSSKEKLQGDAKVVSDAFKKMASQDKDGKTTDADESEKHNYQEQYNKLK
GAGHKKE

>DisProt|DP00592|uniprot|P48539|unigene|Hs.80296|sp|PCP4_HUMAN #1-62

MSERQGAGATNGKDKTSGENDGQKKVQEEFDIDMDAPETERAAVAIQSQFRKFQKKKAGSQS

>DisProt|DP00626|uniprot|P0AG11|sp|UMUD_ECOLI #1-139

MLFIKPADLREIVTFPLFSDLVQCGFPSPAADYVEQRIDLNQLLIQHPSATYFVKASGDSMIDGGI
SDGDLLIVDSAITASHGDIVIAAVDGEFTVKKLQLRPTVQLIPMNSAYSPITISSEDTLDVFGVVIH
VVKAMR

>DisProt|DP00630|uniprot|O76070|unigene|Hs.349470|sp|SYUG_HUMAN #1-127

MDVFKKGFSIAKEGVVGAVEKTKQGVTEAAEKTKEGVMYVGAKTKENVVQSVTSVAEKTKE
QANAVSEAVVSSVNTVATKTVEEAENIAVTSGVVRKEDLRPSAPQQEGEASKEKEEVAEEAQSG
GD

>DisProt|DP00650|uniprot|Q1PAB4|sp|Q1PAB4_9HIV1 #1-101

MEPVDPRLEPWKHPGSQPRTACTNCYCKKCCFHCQVCFIRKALGISYGRKKRRQRRRAPQDSE
THQVSPPKQPASQPRGDPTGPKESKKKVERETETHPVN

>DisProt|DP00665|uniprot|Q9XES8|unigene|Gma.168|sp|Q9XES8_SOYBN #1-89

MAKSKEDITYATSQARLSEDEAVRVAYEHGSPLEGGKIADSQPVDLFSSAHNMPKSGQTTMDS
NTSDQSQMQRDTQEGGSKEFTTGAPG

>DisProt|DP00675_C002|uniprot|P19711|sp|POLG_BVDVN #1-102

SDTKEEGATKKKTQKPDRLERGKMKIVPKESEKDSKTKPPDATIVVEGVKYQVRKKGKTKSKN
TQDGLYHNKNKPQESRKKLEKALLAWAIIAIVLFQVTMG

ANNEX IV

ArchCandy Mutant Dataset

Amyloid-β

>Ab40:

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV

>Ab42:

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA

>Ab43:

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIAT

>A21G: Flemish mutation

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVGGVV

>E22G: Artic mutation

DAEFRHDSGYEVHHQKLVFFAGDVGSNKGAIIGLMVGGVV

>E22Q: Dutch mutation

DAEFRHDSGYEVHHQKLVFFAQDVGSNKGAIIGLMVGGVV

>E22K: Italian mutation

DAEFRHDSGYEVHHQKLVFFAKDVGSNKGAIIGLMVGGVV

>D23N: Iowa mutation

DAEFRHDSGYEVHHQKLVFFAENVGSNKGAIIGLMVGGVV

>T43I: Austrian mutation

DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIAI

>E22del: Japanese mutation

DAEFRHDSGYEVHHQKLVFFADVGSNKGAIIGLMVGGVV

>A21G: Flemish mutation

DAEFRHDSGYEVHHQKLVFFGEDVGSNKGAIIGLMVGGVV

Amylin

>Human Amylin

KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY

>Rat amylin (no fibrils in vivo)

KCNTATCATQRLANFLVRSSNNLGPVLPPTNVGSNTY

>Hamster amylin (no fibrils in vivo)

KCNTATCATQRLANFLVHSNNNFGPVLSPTNVGSNTY

>Degu amylin (no fibrils in vivo)

KCNTATCATQRLTNFLVRSSHNLGAALPPTKVGSNTY

>IPP 8-37

ATCATQRLANFLVHSSNNFGAILSSTNVGSNTY

>IPP 8-37 (3xL):

ATQRLANLLVHSSNNLGAILSSTNVGSNTL

>IPP 8-37 (3xP):

ATQRLANFLPHPSNNFGAILSSPNVGSNTY

>S20G =:

KCNTATCATQRLANFLVHSGNNFGAILSSTNVGSNTY

>S20K =:

KCNTATCATQRLANFLVHSKNNFGAILSSTNVGSNTY

References

Abedini, A. and D. P. Raleigh (2006). "Destabilization of human IAPP amyloid fibrils by proline mutations outside of the putative amyloidogenic domain: is there a critical amyloidogenic domain in human IAPP?" J Mol Biol **355**(2): 274-281.

Ahmed, A. B. and A. V. Kajava (2013). "Breaking the amyloidogenicity code: Methods to predict amyloids from amino acid sequence." FEBS Lett **587**(8): 1089-1095.

Andronesi, O. C., M. von Bergen, J. Biernat, K. Seidel, C. Griesinger, E. Mandelkow and M. Baldus (2008). "Characterization of Alzheimer's-like paired helical filaments from the core domain of tau protein using solid-state NMR spectroscopy." J Am Chem Soc **130**(18): 5922-5928.

Arranz, R., G. Mercado, J. Martín-Benito, R. Giraldo, O. Monasterio, R. Lagos and J. M. Valpuesta (2012). "Structural characterization of microcin E492 amyloid formation: Identification of the precursors." Journal of Structural Biology **178**(1): 54-60.

Astbury, W. T., S. Dickinson and K. Bailey (1935). "The X-ray interpretation of denaturation and the structure of the seed globulins." Biochem J **29**(10): 2351-2360 2351.

Barghorn, S., V. Nimmrich, A. Striebinger, C. Krantz, P. Keller, B. Janson, M. Bahr, M. Schmidt, R. S. Bitner, J. Harlan, E. Barlow, U. Ebert and H. Hillen (2005). "Globular amyloid beta-peptide oligomer - a homogenous and stable neuropathological protein in Alzheimer's disease." J Neurochem **95**(3): 834-847.

Baxa, U., T. Cassese, A. V. Kajava and A. C. Steven (2006). "Structure, function, and amyloidogenesis of fungal prions: filament polymorphism and prion variants." Adv Protein Chem **73**: 125-180.

Baxa, U., T. Cassese, A. V. Kajava and A. C. Steven (2006). Structure, Function, and Amyloidogenesis of Fungal Prions: Filament Polymorphism and Prion Variants. Advances in Protein Chemistry, Elsevier. **73:** 125-180.

Becker, J., N. Ferguson, J. Flinders, B.-J. van Rossum, A. R. Fersht and H. Oschkinat (2008). "A Sequential Assignment Procedure for Proteins that have Intermediate Line Widths in MAS NMR Spectra: Amyloid Fibrils of Human CA150.WW2." ChemBioChem **9**(12): 1946-1952.

Bence, N. F., R. M. Sampat and R. R. Kopito (2001). "Impairment of the ubiquitin-proteasome system by protein aggregation." Science **292**(5521): 1552-1555.

Bennhold, H. (1922). "Specific staining of amyloid by Congo red." MuÈnchener Medizinische Wochenschrift(69): 1537-1538.

Benzinger, T. L., D. M. Gregory, T. S. Burkoth, H. Miller-Auer, D. G. Lynn, R. E. Botto and S. C. Meredith (1998). "Propagating structure of Alzheimer's beta-amyloid(10-35) is parallel beta-sheet with residues in exact register." Proc Natl Acad Sci U S A **95**(23): 13407-13412.

Berson, J. F., D. C. Harper, D. Tenza, G. Raposo and M. S. Marks (2001). "Pmel17 initiates premelanosome morphogenesis within multivesicular bodies." Mol Biol Cell **12**(11): 3451-3464.

Berson, J. F., A. C. Theos, D. C. Harper, D. Tenza, G. Raposo and M. S. Marks (2003). "Proprotein convertase cleavage liberates a fibrillogenic fragment of a resident glycoprotein to initiate melanosome biogenesis." J Cell Biol **161**(3): 521-533.

Bieler, S., L. Estrada, R. Lagos, M. Baeza, J. Castilla and C. Soto (2005). "Amyloid formation modulates the biological activity of a bacterial protein." J Biol Chem **280**(29): 26880-26885.

Bodner, R. A., T. F. Outeiro, S. Altmann, M. M. Maxwell, S. H. Cho, B. T. Hyman, P. J. McLean, A. B. Young, D. E. Housman and A. G. Kazantsev (2006). "Pharmacological promotion of inclusion formation: a therapeutic approach for Huntington's and Parkinson's diseases." Proc Natl Acad Sci U S A **103**(11): 4246-4251.

Boere, H., L. Ruinen and J. H. Scholten (1965). "Electron microscopic studies on the fibrillar component of human splenic amyloid." J Lab Clin Med **66**(6): 943-951.

Bonar, L., A. S. Cohen and M. M. Skinner (1969). "Characterization of the amyloid fibril as a cross-beta protein." Proc Soc Exp Biol Med **131**(4): 1373-1375.

Broersen, K., F. Rousseau and J. Schymkowitz (2010). "The culprit behind amyloid beta peptide related neurotoxicity in Alzheimer's disease: oligomer size or conformation?" Alzheimers Res Ther **2**(4): 12.

Bryan, A. W., Jr., M. Menke, L. J. Cowen, S. L. Lindquist and B. Berger (2009). "BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis." PLoS Comput Biol **5**(3): e1000333.

Cao, P., L. H. Tu, A. Abedini, O. Levsh, R. Akter, V. Patsalo, A. M. Schmidt and D. P. Raleigh (2012). "Sensitivity of amyloid formation by human islet amyloid polypeptide to mutations at residue 20." J Mol Biol **421**(2-3): 282-295.

Carrio, M., N. Gonzalez-Montalban, A. Vera, A. Villaverde and S. Ventura (2005). "Amyloid-like properties of bacterial inclusion bodies." J Mol Biol **347**(5): 1025-1037.

Caughey, B. and P. T. Lansbury (2003). "Protofibrils, pores, fibrils, and neurodegeneration: separating the responsible protein aggregates from the innocent bystanders." Annu Rev Neurosci **26**: 267-298.

Chapman, M. R., L. S. Robinson, J. S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark and S. J. Hultgren (2002). "Role of Escherichia coli curli operons in directing amyloid fiber formation." Science **295**(5556): 851-855.

Chen, A. K. H., R. Y. Y. Lin, E. Z. J. Hsieh, P.-H. Tu, R. P. Y. Chen, T.-Y. Liao, W. Chen, C.-H. Wang and J. J. T. Huang (2010). "Induction of Amyloid Fibrils by the C-Terminal Fragments of TDP-43 in Amyotrophic Lateral Sclerosis." Journal of the American Chemical Society **132**(4): 1186-1187.

Chen, B., K. R. Thurber, F. Shewmaker, R. B. Wickner and R. Tycko (2009). "Measurement of amyloid fibril mass-per-length by tilted-beam transmission electron microscopy." Proc Natl Acad Sci U S A **106**(34): 14339-14344.

Chen, M., M. Margittai, J. Chen and R. Langen (2007). "Investigation of alpha-synuclein fibril structure by site-directed spin labeling." J Biol Chem **282**(34): 24970-24979.

Chien, P., J. S. Weissman and A. H. DePace (2004). "Emerging principles of conformation-based prion inheritance." Annu Rev Biochem **73**: 617-656.

Chiti, F. and C. M. Dobson (2006). "Protein misfolding, functional amyloid, and human disease." Annu Rev Biochem **75**: 333-366.

Chiti, F. and C. M. Dobson (2009). "Amyloid formation by globular proteins under native conditions." Nat Chem Biol **5**(1): 15-22.

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**(6950): 805-808.

Chiti, F., N. Taddei, M. Bucciantini, P. White, G. Ramponi and C. M. Dobson (2000). "Mutational analysis of the propensity for amyloid formation by a globular protein." EMBO J **19**(7): 1441-1449.

Claessen, D., R. Rink, W. de Jong, J. Siebring, P. de Vreugd, F. G. Boersma, L. Dijkhuizen and H. A. Wosten (2003). "A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in Streptomyces coelicolor by forming amyloid-like fibrils." Genes Dev **17**(14): 1714-1726.

Cloe, A. L., J. P. Orgel, J. R. Sachleben, R. Tycko and S. C. Meredith (2011). "The Japanese mutant Abeta (DeltaE22-Abeta(1-39)) forms fibrils instantaneously, with low-thioflavin T fluorescence: seeding of wild-type Abeta(1-40) into atypical fibrils by DeltaE22-Abeta(1-39)." Biochemistry **50**(12): 2026-2039.

Cobb, N. J., F. D. Sönnichsen, H. McHaourab and W. K. Surewicz (2007). "Molecular architecture of human prion protein amyloid: A parallel, in-register β-structure." Proceedings of the National Academy of Sciences **104**(48): 18946-18951.

Cohen, A. S. and E. Calkins (1959). "Electron microscopic observations on a fibrous component in amyloid of diverse origins." Nature **183**(4669): 1202-1203.

Conchillo-Sole, O., N. S. de Groot, F. X. Aviles, 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.

Dalquen, D. A., M. Anisimova, G. H. Gonnet and C. Dessimoz (2012). "ALF--a simulation framework for genome evolution." Mol Biol Evol **29**(4): 1115-1123.

de Groot, N. S., F. X. Aviles, J. Vendrell and S. Ventura (2006). "Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities." FEBS J **273**(3): 658-668.

de Lorenzo, V. (1984). "Isolation and characterization of microcin E492 from Klebsiella pneumoniae." Arch Microbiol **139**(1): 72-75.

Der-Sarkissian, A., C. C. Jao, J. Chen and R. Langen (2003). "Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling." J Biol Chem **278**(39): 37530-37535.

Dobson, C. M. (1999). "Protein misfolding, evolution and disease." Trends Biochem Sci **24**(9): 329-332.

Dos Reis, S. (2001). "The HET-s Prion Protein of the Filamentous Fungus Podospora anserina Aggregates in Vitro into Amyloid-like Fibrils." Journal of Biological Chemistry **277**(8): 5703-5706.

DuBay, K. F., A. P. Pawar, F. Chiti, J. Zurdo, C. M. Dobson and M. Vendruscolo (2004). "Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains." J Mol Biol **341**(5): 1317-1326.

Eanes, E. D. and G. G. Glenner (1968). "X-ray diffraction studies on amyloid filaments." J Histochem Cytochem **16**(11): 673-677.

Elam, J. S., A. B. Taylor, R. Strange, S. Antonyuk, P. A. Doucette, J. A. Rodriguez, S. S. Hasnain, L. J. Hayward, J. S. Valentine, T. O. Yeates and P. J. Hart (2003). "Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS." Nat Struct Biol **10**(6): 461-467.

Esteras-Chopo, A., L. Serrano and M. Lopez de la Paz (2005). "The amyloid stretch hypothesis: recruiting proteins toward the dark side." Proc Natl Acad Sci U S A **102**(46): 16672-16677.

Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper and A. Sali (2007). "Comparative protein structure modeling using MODELLER." Curr Protoc Protein Sci **Chapter 2**: Unit 2 9.

Fandrich, M. (2012). "Oligomeric intermediates in amyloid formation: structure determination and mechanisms of toxicity." J Mol Biol **421**(4-5): 427-440.

Ferguson, N., J. Becker, H. Tidow, S. Tremmel, T. D. Sharpe, G. Krause, J. Flinders, M. Petrovich, J. Berriman, H. Oschkinat and A. R. Fersht (2006). "General structural motifs of amyloid protofilaments." Proc Natl Acad Sci U S A **103**(44): 16248-16253.

Fernandez-Escamilla, A. M., F. Rousseau, J. Schymkowitz and L. Serrano (2004). "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins." Nat Biotechnol **22**(10): 1302-1306.

Fiser, A., R. K. Do and A. Sali (2000). "Modeling of loops in protein structures." Protein Sci **9**(9): 1753-1773.

Fowler, D. M., A. V. Koulov, C. Alory-Jost, M. S. Marks, W. E. Balch and J. W. Kelly (2006). "Functional amyloid formation within mammalian tissue." PLoS Biol **4**(1): e6.

Fox, A., T. Snollaerts, C. Errecart Casanova, A. Calciano, L. A. Nogaj and D. A. Moffet (2010). "Selection for nonamyloidogenic mutants of islet amyloid polypeptide (IAPP) identifies an extended region for amyloidogenicity." Biochemistry **49**(36): 7783-7789.

Fox, A., T. Snollaerts, C. Errecart Casanova, A. Calciano, L. A. Nogaj and D. A. Moffet (2010). "Selection for Nonamyloidogenic Mutants of Islet Amyloid Polypeptide (IAPP) Identifies an Extended Region for Amyloidogenicity." Biochemistry **49**(36): 7783-7789.

Fraser and MacRae (1973). "Conformation in Fibrous Proteins and
Related Synthetic Polypeptides." Academic Press, London and New York.

Friedreich, N., and A. Kekulé (1859). "Zur amyloidfrage." Virchows Archiv **16.1**: 50-65.

Frishman, D. and P. Argos (1995). "Knowledge-based protein secondary structure assignment." Proteins **23**(4): 566-579.

Garbuzynskiy, S. O., M. Y. Lobanov and O. V. Galzitskaya (2010). "FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence." Bioinformatics **26**(3): 326-332.

Gasset, M., M. A. Baldwin, R. J. Fletterick and S. B. Prusiner (1993). "Perturbation of the secondary structure of the scrapie prion protein under conditions that alter infectivity." Proc Natl Acad Sci U S A **90**(1): 1-5.

Geddes, A. J., K. D. Parker, E. D. Atkins and E. Beighton (1968). ""Cross-beta" conformation in proteins." J Mol Biol **32**(2): 343-358.

Gellermann, G. P., T. R. Appel, A. Tannert, A. Radestock, P. Hortschansky, V. Schroeckh, C. Leisner, T. Lutkepohl, S. Shtrasburg, C. Rocken, M. Pras, R. P. Linke, S. Diekmann and M. Fandrich (2005). "Raft lipids as common components of human extracellular amyloid fibrils." Proc Natl Acad Sci U S A **102**(18): 6297-6302.

Giasson, B. I. (2000). "A Hydrophobic Stretch of 12 Amino Acid Residues in the Middle of alpha -Synuclein Is Essential for Filament Assembly." Journal of Biological Chemistry **276**(4): 2380-2386.

Glenner, G. G., D. Ein, E. D. Eanes, H. A. Bladen, W. Terry and D. L. Page (1971). "Creation of "amyloid" fibrils from Bence Jones proteins in vitro." Science **174**(4010): 712-714.

Glenner, G. G. and C. W. Wong (1984). "Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein." Biochem Biophys Res Commun **120**(3): 885-890.

Goldsbury, C., U. Baxa, M. N. Simon, A. C. Steven, A. Engel, J. S. Wall, U. Aebi and S. A. Muller (2011). "Amyloid structure and assembly: insights from scanning transmission electron microscopy." J Struct Biol **173**(1): 1-13.

Goldsbury, C. S., S. Wirtz, S. A. Muller, S. Sunderji, P. Wicki, U. Aebi and P. Frey (2000). "Studies on the in vitro assembly of a beta 1-40: implications for the search for a beta fibril formation inhibitors." J Struct Biol **130**(2-3): 217-231.

Govaerts, C., H. Wille, S. B. Prusiner and F. E. Cohen (2004). "Evidence for assembly of prions with left-handed beta-helices into trimers." Proc Natl Acad Sci U S A **101**(22): 8342-8347.

Greenwald, J. and R. Riek (2010). "Biology of amyloid: structure, function, and regulation." Structure **18**(10): 1244-1260.

Guerois, R., J. E. Nielsen and L. Serrano (2002). "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations." J Mol Biol **320**(2): 369-387.

Guo, Z. and D. Eisenberg (2008). "The structure of a fibril-forming sequence, NNQQNY, in the context of a globular fold." Protein Sci **17**(9): 1617-1623.

Hamodrakas, S. J., C. Liappa and V. A. Iconomidou (2007). "Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins." Int J Biol Macromol **41**(3): 295-300.

Harper, J. D. and P. T. Lansbury, Jr. (1997). "Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins." Annu Rev Biochem **66**: 385-407.

Hearing, V. J. (2000). "The melanosome: the perfect model for cellular responses to the environment." Pigment Cell Res **13 Suppl 8**: 23-34.

Hebert, L. E., P. A. Scherr, J. L. Bienias, D. A. Bennett and D. A. Evans (2003). "Alzheimer disease in the US population: prevalence estimates using the 2000 census." Arch Neurol **60**(8): 1119-1122.

Hennetin, J., B. Jullian, A. C. Steven and A. V. Kajava (2006). "Standard conformations of beta-arches in beta-solenoid proteins." J Mol Biol **358**(4): 1094-1105.

Hy, L. X. and D. M. Keller (2000). "Prevalence of AD among whites: a summary by levels of severity." Neurology **55**(2): 198-204.

Iconomidou, V. A. and S. J. Hamodrakas (2008). "Natural protective amyloids." Curr Protein Pept Sci **9**(3): 291-309.

Inouye, H. and D. A. Kirschner (2006). "X-Ray fiber and powder diffraction of PrP prion peptides." Adv Protein Chem **73**: 181-215.

Iwata, K., T. Fujiwara, Y. Matsuki, H. Akutsu, S. Takahashi, H. Naiki and Y. Goto (2006). "3D structure of amyloid protofilaments of beta2-microglobulin fragment probed by solid-state NMR." Proc Natl Acad Sci U S A **103**(48): 18119-18124.

Iwata, K., T. Fujiwara, Y. Matsuki, H. Akutsu, S. Takahashi, H. Naiki and Y. Goto (2006). "3D structure of amyloid protofilaments of β2-microglobulin fragment probed by solid-state NMR." Proceedings of the National Academy of Sciences **103**(48): 18119-18124.

Jaroniec, C. P., C. E. MacPhee, V. S. Bajaj, M. T. McMahon, C. M. Dobson and R. G. Griffin (2004). "High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy." Proc Natl Acad Sci U S A **101**(3): 711-716.

Jeganathan, S., M. von Bergen, E. M. Mandelkow and E. Mandelkow (2008). "The natively unfolded character of tau and its aggregation to Alzheimer-like paired helical filaments." Biochemistry **47**(40): 10526-10539.

Jimenez, J. L., J. I. Guijarro, E. Orlova, J. Zurdo, C. M. Dobson, M. Sunde and H. R. Saibil (1999). "Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing." EMBO J **18**(4): 815-821.

Jimenez, J. L., E. J. Nettleton, M. Bouchard, C. V. Robinson, C. M. Dobson and H. R. Saibil (2002). "The protofilament structure of insulin amyloid fibrils." Proc Natl Acad Sci U S A **99**(14): 9196-9201.

Kajava, A. V. (2012). "Tandem repeats in proteins: from sequence to structure." J Struct Biol **179**(3): 279-288.
Kajava, A. V., U. Aebi and A. C. Steven (2005). "The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin." J Mol Biol **348**(2): 247-252.
Kajava, A. V., U. Baxa and A. C. Steven (2010). "Beta arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils." FASEB J **24**(5): 1311-1319.
Kajava, A. V., U. Baxa, R. B. Wickner and A. C. Steven (2004). "A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure." Proc Natl Acad Sci U S A **101**(21): 7885-7890.
Kajava, A. V. and A. C. Steven (2006). "Beta-rolls, beta-helices, and other beta-solenoid proteins." Adv Protein Chem **73**: 55-96.
Kamihira, M., A. Naito, S. Tuzi, A. Y. Nosaka and H. Saito (2000). "Conformational transitions and fibrillation mechanism of human calcitonin as studied by high-resolution solid-state 13C NMR." Protein Science **9**(05): 867-877.
Kelly, R. B. (1985). "Pathways of protein secretion in eukaryotes." Science **230**(4721): 25-32.
Kim, C., J. Choi, S. J. Lee, W. J. Welsh and S. Yoon (2009). "NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation." Nucleic Acids Res **37**(Web Server issue): W469-473.
Kirschner, D. A., C. Abraham and D. J. Selkoe (1986). "X-ray diffraction from intraneuronal paired helical filaments and extraneuronal amyloid fibers in Alzheimer disease indicates cross-beta conformation." Proceedings of the National Academy of Sciences of the United States of America **83**(2).
Kopito, R. R. (2000). "Aggresomes, inclusion bodies and protein aggregation." Trends Cell Biol **10**(12): 524-530.
Kretzschmar, H. and J. Tatzelt (2013). "Prion disease: a tale of folds and strains." Brain Pathol **23**(3): 321-332.
Krishnan, R. and S. L. Lindquist (2005). "Structural insights into a yeast prion illuminate nucleation and strain diversity." Nature **435**(7043): 765-772.
Lashuel, H. A., D. Hartley, B. M. Petre, T. Walz and P. T. Lansbury, Jr. (2002). "Neurodegenerative disease: amyloid pores from pathogenic mutations." Nature **418**(6895): 291.
Lashuel, H. A. and P. T. Lansbury, Jr. (2006). "Are amyloid diseases caused by protein aggregates that mimic bacterial pore-forming toxins?" Q Rev Biophys **39**(2): 167-201.
Li, J., T. McQuade, Ansgar B. Siemer, J. Napetschnig, K. Moriwaki, Y.-S. Hsiao, E. Damko, D. Moquin, T. Walz, A. McDermott, Francis K.-M. Chan and H. Wu (2012). "The RIP1/RIP3 Necrosome Forms a Functional Amyloid Signaling Complex Required for Programmed Necrosis." Cell **150**(2): 339-350.
Liu, Y. and B. Kuhlman (2006). "RosettaDesign server for protein design." Nucleic Acids Res **34**(Web Server issue): W235-238.
Lopez de la Paz, M. and L. Serrano (2004). "Sequence determinants of amyloid fibril formation." Proc Natl Acad Sci U S A **101**(1): 87-92.
Lotz, G. P. and J. Legleiter (2013). "The role of amyloidogenic protein oligomerization in neurodegenerative disease." J Mol Med (Berl).
Luca, S., W. M. Yau, R. Leapman and R. Tycko (2007). "Peptide conformation and supramolecular organization in amylin fibrils: constraints from solid-state NMR." Biochemistry **46**(47): 13505-13522.

Luhrs, T., C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Dobeli, D. Schubert and R. Riek (2005). "3D structure of Alzheimer's amyloid-beta(1-42) fibrils." Proc Natl Acad Sci U S A **102**(48): 17342-17347.

Ma, B. and R. Nussinov (2002). "Stabilities and conformations of Alzheimer's beta - amyloid peptide oligomers (Abeta 16-22, Abeta 16-35, and Abeta 10-35): Sequence effects." Proc Natl Acad Sci U S A **99**(22): 14126-14131.

Maji, S. K., M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale and R. Riek (2009). "Functional amyloids as natural storage of peptide hormones in pituitary secretory granules." Science **325**(5938): 328-332.

Marcelino-Cruz, A. M., M. Bhattacharya, A. C. Anselmo and P. M. Tessier (2011). "Site-specific structural analysis of a yeast prion strain with species-specific seeding activity." Prion **5**(3): 208-214.

Marek, P., A. Abedini, B. Song, M. Kanungo, M. E. Johnson, R. Gupta, W. Zaman, S. S. Wong and D. P. Raleigh (2007). "Aromatic interactions are not required for amyloid fibril formation by islet amyloid polypeptide but do influence the rate of fibril formation and fibril morphology." Biochemistry **46**(11): 3255-3261.

Marek, P., S. Mukherjee, M. T. Zanni and D. P. Raleigh (2010). "Residue-specific, real-time characterization of lag-phase species and fibril growth during amyloid formation: a combined fluorescence and IR study of p-cyanophenylalanine analogs of islet amyloid polypeptide." J Mol Biol **400**(4): 878-888.

Margittai, M. and R. Langen (2004). "Template-assisted filament growth by parallel stacking of tau." Proc Natl Acad Sci U S A **101**(28): 10278-10283.

Margittai, M. and R. Langen (2008). "Fibrils with parallel in-register structure constitute a major class of amyloid fibrils: molecular insights from electron paramagnetic resonance spectroscopy." Q Rev Biophys **41**(3-4): 265-297.

Marston, F. A. (1986). "The purification of eukaryotic polypeptides synthesized in Escherichia coli." Biochem J **240**(1): 1-12.

Maurer-Stroh, S., M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, I. C. Martins, J. Reumers, K. L. Morris, A. Copland, L. Serpell, L. Serrano, J. W. Schymkowitz and F. Rousseau (2010). "Exploring the sequence determinants of amyloid structure using position-specific scoring matrices." Nat Methods **7**(3): 237-242.

Meinhardt, J., C. Sachse, P. Hortschansky, N. Grigorieff and M. Fandrich (2009). "Abeta(1-40) fibril polymorphism implies diverse interaction patterns in amyloid fibrils." J Mol Biol **386**(3): 869-877.

Miravalle, L., T. Tokuda, R. Chiarle, G. Giaccone, O. Bugiani, F. Tagliavini, B. Frangione and J. Ghiso (2000). "Substitutions at codon 22 of Alzheimer's abeta peptide induce diverse conformational changes and apoptotic effects in human cerebral endothelial cells." J Biol Chem **275**(35): 27110-27116.

Murakami, K., K. Irie, A. Morimoto, H. Ohigashi, M. Shindo, M. Nagao, T. Shimizu and T. Shirasawa (2003). "Neurotoxicity and physicochemical properties of Abeta mutant peptides from cerebral amyloid angiopathy: implication for the pathogenesis of cerebral amyloid angiopathy and Alzheimer's disease." J Biol Chem **278**(46): 46179-46187.

Nelson, R. and D. Eisenberg (2006). "Structural models of amyloid-like fibrils." Adv Protein Chem **73**: 235-282.

Nelson, R., M. R. Sawaya, M. Balbirnie, A. O. Madsen, C. Riekel, R. Grothe and D. Eisenberg (2005). "Structure of the cross-beta spine of amyloid-like fibrils." Nature **435**(7043): 773-778.

Nilsberth, C., J. Luthman, L. Lannfelt and M. Schultzberg (1999). "Expression of presenilin 1 mRNA in rat peripheral organs and brain." Histochem J **31**(8): 515-523.
Olsen, A., A. Arnqvist, M. Hammar and S. Normark (1993). "Environmental regulation of curli production in Escherichia coli." Infect Agents Dis **2**(4): 272-274.
Osherovich, L. Z., B. S. Cox, M. F. Tuite and J. S. Weissman (2004). "Dissection and design of yeast prions." PLoS Biol **2**(4): E86.
Pal, C., B. Papp and M. J. Lercher (2006). "An integrated view of protein evolution." Nat Rev Genet **7**(5): 337-348.
Paravastu, A. K., R. D. Leapman, W. M. Yau and R. Tycko (2008). "Molecular structural basis for polymorphism in Alzheimer's beta-amyloid fibrils." Proc Natl Acad Sci U S A **105**(47): 18349-18354.
Parvathy, P. D., Vahram Haroutunian, Dushyant P. Purohit, Kenneth L. Davis, Richard C. Mohs, Helen Park; Thomas M. Moran, Joseph Y. Chan, Joseph D. Buxbaum, (2001). "Correlation Between Aβx-40–, Aβx-42–, and Aβx-43–Containing Amyloid Plaques and Cognitive Decline." JAMA Neurology **58**.
Paul Emsley, B. L., William G. Scott, Kevin Cowtan (2010). "Features and Development of Coot." Acta Crystallographica Section D - Biological Crystallography **66**: 486-501.
Pepys, M. B. (2006). "Amyloidosis." Annu Rev Med **57**: 223-241.
Perutz, M. F., B. J. Pope, D. Owen, E. E. Wanker and E. Scherzinger (2002). "Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid β-peptide of amyloid plaques." Proceedings of the National Academy of Sciences **99**(8): 5596-5600.
Petkova, A. T., W. M. Yau and R. Tycko (2006). "Experimental constraints on quaternary structure in Alzheimer's beta-amyloid fibrils." Biochemistry **45**(2): 498-512.
Prusiner, S. B. (1982). "Novel proteinaceous infectious particles cause scrapie." Science **216**(4542): 136-144.
Qiang, W., W. M. Yau, Y. Luo, M. P. Mattson and R. Tycko (2012). "Antiparallel beta-sheet architecture in Iowa-mutant beta-amyloid fibrils." Proc Natl Acad Sci U S A **109**(12): 4443-4448.
Riley, M. A. (1998). "Molecular mechanisms of bacteriocin evolution." Annu Rev Genet **32**: 255-278.
Ritter, C., M. L. Maddelein, A. B. Siemer, T. Luhrs, M. Ernst, B. H. Meier, S. J. Saupe and R. Riek (2005). "Correlation of structural elements and infectivity of the HET-s prion." Nature **435**(7043): 844-848.
Rojas Quijano, F. A., D. Morrow, B. M. Wise, F. L. Brancia and W. J. Goux (2006). "Prediction of nucleating sequences from amyloidogenic propensities of tau-related peptides." Biochemistry **45**(14): 4638-4652.
Sachse, C., M. Fandrich and N. Grigorieff (2008). "Paired beta-sheet structure of an Abeta(1-40) amyloid fibril revealed by electron microscopy." Proc Natl Acad Sci U S A **105**(21): 7462-7466.
Sakagashira, S., H. J. Hiddinga, K. Tateishi, T. Sanke, T. Hanabusa, K. Nanjo and N. L. Eberhardt (2000). "S20G mutant amylin exhibits increased in vitro amyloidogenicity and increased intracellular cytotoxicity compared to wild-type amylin." Am J Pathol **157**(6): 2101-2109.
Sakagashira, S., T. Sanke, T. Hanabusa, H. Shimomura, S. Ohagi, K. Y. Kumagaye, K. Nakajima and K. Nanjo (1996). "Missense mutation of amylin gene (S20G) in Japanese NIDDM patients." Diabetes **45**(9): 1279-1281.

Sanders, A., C. Jeremy Craven, L. D. Higgins, S. Giannini, M. J. Conroy, A. M. Hounslow, J. P. Waltho and R. A. Staniforth (2004). "Cystatin forms a tetramer through structural rearrangement of domain-swapped dimers prior to amyloidogenesis." J Mol Biol **336**(1): 165-178.

Santoso, A., P. Chien, L. Z. Osherovich and J. S. Weissman (2000). "Molecular Basis of a Yeast Prion Species Barrier." Cell **100**(2): 277-288.

Saveanu, L., D. Fruci and P. van Endert (2002). "Beyond the proteasome: trimming, degradation and generation of MHC class I ligands by auxiliary proteases." Mol Immunol **39**(3-4): 203-215.

Sawaya, M. R., S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. Wiltzius, H. T. McFarlane, A. O. Madsen, C. Riekel and D. Eisenberg (2007). "Atomic structures of amyloid cross-beta spines reveal varied steric zippers." Nature **447**(7143): 453-457.

Sawyer, E. B., D. Claessen, M. Haas, B. Hurgobin and S. L. Gras (2011). "The Assembly of Individual Chaplin Peptides from Streptomyces coelicolor into Functional Amyloid Fibrils." PLoS ONE **6**(4).

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Sen, A., U. Baxa, M. N. Simon, J. S. Wall, R. Sabate, S. J. Saupe and A. C. Steven (2007). "Mass analysis by scanning transmission electron microscopy and electron diffraction validate predictions of stacked beta-solenoid model of HET-s prion fibrils." J Biol Chem **282**(8): 5545-5550.

Serag, A. A., C. Altenbach, M. Gingery, W. L. Hubbell and T. O. Yeates (2002). "Arrangement of subunits and ordering of beta-strands in an amyloid sheet." Nat Struct Biol **9**(10): 734-739.

Sharma, D., L. M. Shinchuk, H. Inouye, R. Wetzel and D. A. Kirschner (2005). "Polyglutamine homopolymers having 8-45 residues form slablike beta-crystallite assemblies." Proteins **61**(2): 398-411.

Shewmaker, F., R. P. McGlinchey, K. R. Thurber, P. McPhie, F. Dyda, R. Tycko and R. B. Wickner (2009). "The Functional Curli Amyloid Is Not Based on In-register Parallel -Sheet Structure." Journal of Biological Chemistry **284**(37): 25065-25076.

Shewmaker, F., R. P. McGlinchey, K. R. Thurber, P. McPhie, F. Dyda, R. Tycko and R. B. Wickner (2009). "The functional curli amyloid is not based on in-register parallel beta-sheet structure." J Biol Chem **284**(37): 25065-25076.

Shewmaker, F., R. B. Wickner and R. Tycko (2006). "Amyloid of the prion domain of Sup35p has an in-register parallel beta-sheet structure." Proc Natl Acad Sci U S A **103**(52): 19754-19759.

Shirahama, T. and A. S. Cohen (1965). "Structure of amyloid fibrils after negative staining and high-resolution electron microscopy." Nature **206**(985): 737-738.

Si, K., Y. B. Choi, E. White-Grindley, A. Majumdar and E. R. Kandel (2010). "Aplysia CPEB can form prion-like multimers in sensory neurons that contribute to long-term facilitation." Cell **140**(3): 421-435.

Si, K., M. Giustetto, A. Etkin, R. Hsu, A. M. Janisiewicz, M. C. Miniaci, J. H. Kim, H. Zhu and E. R. Kandel (2003). "A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in aplysia." Cell **115**(7): 893-904.

Siemer, A. B., C. Ritter, M. Ernst, R. Riek and B. H. Meier (2005). "High-resolution solid-state NMR spectroscopy of the prion protein HET-s in its amyloid conformation." Angew Chem Int Ed Engl **44**(16): 2441-2444.

Sikorski, P. and E. Atkins (2005). "New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils." Biomacromolecules **6**(1): 425-432.

Simons, K. T., R. Bonneau, I. Ruczinski and D. Baker (1999). "Ab initio protein structure prediction of CASP III targets using ROSETTA." Proteins **Suppl 3**: 171-176.
Sparr, E., M. F. Engel, D. V. Sakharov, M. Sprong, J. Jacobs, B. de Kruijff, J. W. Hoppener and J. A. Killian (2004). "Islet amyloid polypeptide-induced membrane leakage involves uptake of lipids by forming amyloid fibers." FEBS Lett **577**(1-2): 117-120.
Stefani, M. and C. M. Dobson (2003). "Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution." J Mol Med (Berl) **81**(11): 678-699.
Sunde, M. and C. Blake (1997). "The structure of amyloid fibrils by electron microscopy and X-ray diffraction." Adv Protein Chem **50**: 123-159.
Sunde, M., L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys and C. C. Blake (1997). "Common core structure of amyloid fibrils by synchrotron X-ray diffraction." J Mol Biol **273**(3): 729-739.
Symmers, W. S. (1956). "Primary amyloidosis: a review." J Clin Pathol **9**(3): 187-211.
Tartaglia, G. G., A. Cavalli and M. Vendruscolo (2007). "Prediction of local structural stabilities of proteins from their amino acid sequences." Structure **15**(2): 139-143.
Tartaglia, G. G., R. Pellarin, A. Cavalli and A. Caflisch (2005). "Organism complexity anti-correlates with proteomic beta-aggregation propensity." Protein Sci **14**(10): 2735-2740.
Tartaglia, G. G. and M. Vendruscolo (2008). "The Zyggregator method for predicting protein aggregation propensities." Chem Soc Rev **37**(7): 1395-1401.
Termine, J. D., E. D. Eanes, D. Ein and G. G. Glenner (1972). "Infrared spectroscopy of human amyloid fibrils and immunoglobulin proteins." Biopolymers **11**(5): 1103-1113.
Thakur, A. K. and R. Wetzel (2002). "Mutational analysis of the structural organization of polyglutamine aggregates." Proc Natl Acad Sci U S A **99**(26): 17014-17019.
Thompson, M. J., S. A. Sievers, J. Karanicolas, M. I. Ivanova, D. Baker and D. Eisenberg (2006). "The 3D profile method for identifying fibril-forming segments of proteins." Proc Natl Acad Sci U S A **103**(11): 4074-4078.
Torok, M., S. Milton, R. Kayed, P. Wu, T. McIntire, C. G. Glabe and R. Langen (2002). "Structural and dynamic features of Alzheimer's Abeta peptide in amyloid fibrils studied by site-directed spin labeling." J Biol Chem **277**(43): 40810-40815.
Toyama, B. H. and J. S. Weissman (2011). "Amyloid structure: conformational diversity and consequences." Annu Rev Biochem **80**: 557-585.
Trovato, A., F. Seno and S. C. Tosatto (2007). "The PASTA server for protein aggregation prediction." Protein Eng Des Sel **20**(10): 521-523.
True, H. L. and S. L. Lindquist (2000). "A yeast prion provides a mechanism for genetic variation and phenotypic diversity." Nature **407**(6803): 477-483.
Tycko, R. (2011). "Solid-state NMR studies of amyloid fibril structure." Annu Rev Phys Chem **62**: 279-299.
Van Der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. Berendsen (2005). "GROMACS: fast, flexible, and free." J Comput Chem **26**(16): 1701-1718.
Van Melckebeke, H., C. Wasmer, A. Lange, E. Ab, A. Loquet, A. Bockmann and B. H. Meier (2010). "Atomic-resolution three-dimensional structure of HET-s(218-289) amyloid fibrils by solid-state NMR spectroscopy." J Am Chem Soc **132**(39): 13765-13775.
Van Nostrand, W. E., J. P. Melchor, H. S. Cho, S. M. Greenberg and G. W. Rebeck (2001). "Pathogenic effects of D23N Iowa mutant amyloid beta -protein." J Biol Chem **276**(35): 32860-32866.

Vidal, O., R. Longin, C. Prigent-Combaret, C. Dorel, M. Hooreman and P. Lejeune (1998). "Isolation of an Escherichia coli K-12 mutant strain able to form biofilms on inert surfaces: involvement of a new ompR allele that increases curli expression." J Bacteriol **180**(9): 2442-2449.

Virchow, R. (1854). "Virchows Arch. f. pathol. ." Anat. u. Physiol: 3.139.

Vucetic, S., Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton and A. K. Dunker (2005). "DisProt: a database of protein disorder." Bioinformatics **21**(1): 137-140.

Wang, J., S. Gulich, C. Bradford, M. Ramirez-Alvarado and L. Regan (2005). "A twisted four-sheeted model for an amyloid fibril." Structure **13**(9): 1279-1288.

Wasmer, C., A. Lange, H. Van Melckebeke, A. B. Siemer, R. Riek and B. H. Meier (2008). "Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core." Science **319**(5869): 1523-1526.

Watt, B., G. van Niel, D. M. Fowler, I. Hurbain, K. C. Luk, S. E. Stayrook, M. A. Lemmon, G. Raposo, J. Shorter, J. W. Kelly and M. S. Marks (2009). "N-terminal Domains Elicit Formation of Functional Pmel17 Amyloid Fibrils." Journal of Biological Chemistry **284**(51): 35543-35555.

Wessels, J. G. (1997). "Hydrophobins: proteins that change the nature of the fungal surface." Adv Microb Physiol **38**: 1-45.

Westermark, P., K. H. Johnson, T. D. O'Brien and C. Betsholtz (1992). "Islet amyloid polypeptide--a novel controversy in diabetes research." Diabetologia **35**(4): 297-303.

Westermark, P., K. Sletten, B. Johansson and G. G. Cornwell, 3rd (1990). "Fibril in senile systemic amyloidosis is derived from normal transthyretin." Proc Natl Acad Sci U S A **87**(7): 2843-2845.

William E. Van Nostrand, J. P. M., Hyun Soon Cho, Steven M. Greenberg, and a. G. W. Rebeck (2001). "Pathogenic Effects of D23N Iowa Mutant Amyloid Beta-Protein." JBC

Williams, A. D., E. Portelius, I. Kheterpal, J. T. Guo, K. D. Cook, Y. Xu and R. Wetzel (2004). "Mapping abeta amyloid fibril secondary structure using scanning proline mutagenesis." J Mol Biol **335**(3): 833-842.

Wiltzius, J. J., S. A. Sievers, M. R. Sawaya, D. Cascio, D. Popov, C. Riekel and D. Eisenberg (2008). "Atomic structure of the cross-beta spine of islet amyloid polypeptide (amylin)." Protein Sci **17**(9): 1467-1474.

Wosten, H. A. and M. L. de Vocht (2000). "Hydrophobins, the fungal coat unravelled." Biochim Biophys Acta **1469**(2): 79-86.

Wosten, H. A. and J. M. Willey (2000). "Surface-active proteins enable microbial aerial hyphae to grow into the air." Microbiology **146 (Pt 4)**: 767-773.

Wurth, C., N. K. Guimard and M. H. Hecht (2002). "Mutations that reduce aggregation of the Alzheimer's Abeta42 peptide: an unbiased search for the sequence determinants of Abeta amyloidogenesis." J Mol Biol **319**(5): 1279-1290.

Ye, Z., D. Bayron Poueymiroy, J. J. Aguilera, S. Srinivasan, Y. Wang, L. C. Serpell and W. Colón (2011). "Inflammation Protein SAA2.2 Spontaneously Forms Marginally Stable Amyloid Fibrils at Physiological Temperature." Biochemistry **50**(43): 9184-9191.

Ye, Z., K. C. French, L. A. Popova, I. K. Lednev, M. M. Lopez and G. I. Makhatadze (2009). "Mechanism of Fibril Formation by a 39-Residue Peptide (PAPf39) from Human Prostatic Acidic Phosphatase." Biochemistry **48**(48): 11582-11591.

Yin, Y. I., B. Bassit, L. Zhu, X. Yang, C. Wang and Y. M. Li (2007). "{gamma}-Secretase Substrate Concentration Modulates the Abeta42/Abeta40 Ratio:

IMPLICATIONS FOR ALZHEIMER DISEASE." J Biol Chem **282**(32): 23639-23644.

Yoon, S. and W. J. Welsh (2004). "Detecting hidden sequence propensity for amyloid fibril formation." Protein Sci **13**(8): 2149-2160.

Zhang-Nunes, S. X., M. L. Maat-Schieman, S. G. van Duinen, R. A. Roos, M. P. Frosch and S. M. Greenberg (2006). "The cerebral beta-amyloid angiopathies: hereditary and sporadic." Brain Pathol **16**(1): 30-39.

Zhang, Z., H. Chen and L. Lai (2007). "Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential." Bioinformatics **23**(17): 2218-2225.