Biophysical Reviews (2021) 13:71–89
https://doi.org/10.1007/s12551-021-00778-w

REVIEW

Protein aggregation: in silico algorithms and applications

R. Prabakaran[1] · Puneet Rawat[1] · A. Mary Thangakani[1] · Sandeep Kumar[2] · M. Michael Gromiha[1,3]

Received: 6 November 2020 / Accepted: 1 January 2021 / Published online: 17 January 2021

Abstract

Protein aggregation is a topic of immense interest to the scientific community due to its role in several neurodegenerative diseases/disorders and industrial importance. Several in silico techniques, tools, and algorithms have been developed to predict aggregation in proteins and understand the aggregation mechanisms. This review attempts to provide an essence of the vast developments in in silico approaches, resources available, and future perspectives. It reviews aggregation-related databases, mechanistic models (aggregation-prone region and aggregation propensity prediction), kinetic models (aggregation rate prediction), and molecular dynamics studies related to aggregation. With a multitude of prediction models related to aggregation already available to the scientific community, the field of protein aggregation is rapidly maturing to tackle new applications.

Keywords Protein aggregation · Peptide assembly · Aggregation kinetics · Aggregation propensity · Prediction · Algorithm · Molecular dynamics

Introduction

The human proteome consists of more than 20,000 proteins with diverse sizes, compositions, structures, and functions. Almost every cellular activity depends on the conformation and concentration of proteins. Normal cellular processes are tightly aligned with proteostasis, which involves synthesis, folding, trafficking, and degradation of the protein. Internal or external perturbations that disturb the proteostasis could lead to loss of protein functions, to changes in protein turnover rate and protein concentration, and potentially to undesirable consequences such as deposition of protein aggregates in the affected tissues and organs. Protein aggregation, which can be fibrillar or amorphous, has been studied over several decades (Astbury et al. 1935; Green and Hughes 1955; Kyle and Bayrd 1975) from both physiochemical and pathological perspectives. The recent surge of interest in protein aggregation is attributed to two crucial applications: the association of cross β-steric zipper–rich aggregates in human proteinopathies and the development of protein-based therapeutic molecules.

Protein and peptide therapeutics are promising classes of medicines with vast and growing clinical applications. Protein therapeutics include monoclonal antibodies (mAb), hormones, vaccines, enzymes, growth factors, fusion proteins, and so on (Leader et al. 2008; Kintzing et al. 2016; Lagassé et al. 2017; Usmani et al. 2017). The manufacturing of these biotherapeutics is tedious and often complicated by protein instability and aggregation. As a result, developability assessments and optimization to increase protein stability and solubility as well as decrease viscosity and aggregation have become a critical step in biotherapeutic drug discovery and development (Wang et al. 2009; Zurdo 2013; Li et al. 2016; Jain et al. 2017). The aggregation propensity of a biopharmaceutical can potentially affect its solubility and the viscosity of its liquid formulations. Low solubility and high viscosity often translate to difficulties in drug delivery, manufacturing, and storage (Roberts 2014). In order to assess and improve the developability, several in silico approaches have been developed for designing and optimizing therapeutic proteins and peptides (Nichols et al. 2015; Agrawal et al. 2016).

Neurodegenerative diseases such as Parkinson's, Alzheimer's, prion, etc. are characterized by progressive nervous system dysfunction (Iadanza et al. 2018). Despite diverse

✉ Sandeep Kumar
Sandeep_2.Kumar@Boehringer-Ingelheim.com

✉ M. Michael Gromiha
gromiha@iitm.ac.in

[1] Department of Biotechnology, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

[2] Biotherapeutics Discovery, Boehringer Ingelheim Pharmaceutical Inc., Ridgefield, CT, USA

[3] School of Computing, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Kanagawa, Japan

risk factors such as ageing, environmental factors, and genetic mutations, the accumulation of intracellular and extracellular proteinaceous deposits is considered the key factor. Amyloidosis refers to a group of diseases associated with the deposition of amyloid fibrils leading to pathogenesis (Benson et al. 2018; Ke et al. 2020). However, the word 'amyloid' was derived from the Latin word 'amylum' for starch, and the amyloid deposits are made up of long, 50- to 200-Å-wide, β-sheet–rich protein fibrils (Kyle and Bayrd 1975). The characteristics and location of these deposits and the symptoms associated with the disease vary depending on the protein involved in the amyloidosis (Gertz 2018). For example, amyloid light-chain (AL) amyloidosis, one of the common amyloidosis, consists of deposition of immunoglobulin light chains in the kidney and heart (Dogan 2017).

Apart from the role in human pathology and ageing, protein aggregates have been found to also play functional roles in several organisms: For example, Curlin in *E. coli* to mediate host interaction and Ure2p in *Saccharomyces cerevisiae* to regulate nitrogen intake (Chiti and Dobson 2006). The ability of a protein to form amyloid fibrils is attributed to the aggregation-prone region (APR) in the protein sequence (Ventura et al. 2004; Esteras-Chopo et al. 2005). These APRs mediate intermolecular self-interactions leading to cross β-steric zipper formation, which forms the stable core of fibrillar macromolecular structures found in amyloid deposits (Nelson et al. 2005; Sawaya et al. 2007).

Databases for protein aggregation

The exponential increase in the experimental data related to protein aggregation in the last few years has led to the necessity of storing and curating the information related to protein aggregation. Currently, there are several databases available to assist the scientific community (Table 1). These specialized protein aggregation–related databases contain comprehensive, extended knowledge from literature. Fibril_one (Siepen and Westhead 2002) was the first amyloidogenic protein database containing 250 mutations and 50 experimental conditions associated with 22 proteins. Lopez de la Paz and Serrano (2004) curated the amyloidogenic peptides by systematically mutating the residues of amyloidogenic STVIIE peptide. The dataset was extended with the inclusion of peptides from insulin, β2-microglobulin, amylin, tau protein, etc. (Thompson et al. 2006). Goldschmidt et al. (2010) predicted the aggregation profile of 76 genomes and created the ZipperDB database. WALTZ-DB (Beerten et al. 2015) is a collection for experimentally known amyloid-forming hexapeptides, characterized using electron microscopy, dye binding, and Fourier transform infrared spectroscopy. WALTZ-DB was recently updated to WALTZ-DB 2.0 (Louros et al. 2020) by expanding the hexapeptide sequence dataset and adding new structural information. Angarica et al. (2014) developed the database PrionScan for predicted prion-like domains in complete proteomes. Around the same time, Shobana and Pandaranayaka (2014) constructed the integrated database ProADD for the diseases caused by protein aggregation along with the proteins involved in aggregation. The AmyLoad (Wozniak and Kotulska 2015) database compiled amyloidogenic and non-amyloidogenic sequence fragments from various sources (Conchillo-Solé et al. 2007; Fernandez-Escamilla et al. 2004; Goldschmidt et al. 2010) as well as from literature. AmyPro (Varadi et al. 2018) is a recently developed comprehensive database on precursor proteins and their aggregation-prone regions.

Thangakani et al. (2016) developed the comprehensive database CPAD on experimentally verified aggregating proteins, aggregation-prone regions of different lengths, and aggregation kinetics. CPAD has been updated recently to CPAD 2.0 and includes a new category of aggregation-related structures (Rawat et al. 2020a). AmyloBase (Belli et al. 2011) was the first resource for the aggregation kinetics experiments. AMYPdb (Pawlicki et al. 2008) curates the structural information of amyloidogenic proteins and currently has 1200 structures from 31 amyloidogenic protein families. The protein families are clustered based on amyloidogenic sequence pattern. PDB_Amyloid (Takács et al. 2019) is a recently developed database containing a list of amyloid structures and globular structure entries with an amyloid-like substructure. AL-Base (Bodi et al. 2009) is a curated database of light-chain sequence of antibodies derived from patients with light-chain (AL) amyloidosis.

In silico methods and tools for protein aggregation

Over the last few decades, several computational techniques and tools have been developed to address protein aggregation. These tools and techniques can be broadly classified into three classes: (a) APR and aggregation propensity prediction, (b) aggregation kinetics prediction, and (c) molecular simulation techniques.

Aggregation-prone region and aggregation propensity prediction

Amyloid fibrils are composed of a cross-β steric zipper motif (Fig. 1), which consists of stacked β-strands with interdigitating side chains and axially oriented backbone hydrogen bonds (Sunde and Blake 1997). Nucleation of the cross-beta steric zipper motif (Sawaya et al. 2007) and its assembly into amyloid fibrils depend on several extrinsic factors such as temperature, pH, and protein and ionic concentration as well as on the intrinsic ones such as amino acid composition and sequence

Table 1 Protein aggregation databases

Database	Link	Reference
FIbril_one*	http://www.bioinformatics.leeds.ac.uk/group/online/fibril_one	Siepen and Westhead (2002)
ZipperDB	https://services.mbi.ucla.edu/zipperdb/	Goldschmidt et al. (2010)
WALTZ-DB 2.0	http://waltzdb.switchlab.org/	Louros et al. (2020)
PrionScan	http://webapps.bifi.es/prionscan	Angarica et al. (2014)
ProADD*	http://bicmku.in/ProADD	Shobana and Pandaranayaka (2014)
AmyLoad	http://comprec-lin.iiar.pwr.edu.pl/amyload/	Wozniak and Kotulska (2015)
AmyPro	https://amypro.net/#/	Varadi et al. (2018)
CPAD 2.0	https://web.iitm.ac.in/bioinfo2/cpad2/	Rawat et al. (2020a)
AmyloBase	http://150.217.63.173/biochimica/bioinfo/amylobase/pages/view.html	Belli et al. (2011)
AMYPdb	http://amypdb.genouest.org/e107_plugins/amypdb_project/project.php	Pawlicki et al. (2008)
PDB_Amyloid	https://pitgroup.org/amyloid/	Takács et al. (2019)
AL-Base	http://albase.bumc.bu.edu/aldb	Bodi et al. (2009)

*The database is no longer available online as per last access on 1 October 2020

patterning of short aggregation-prone regions (APRs), which are typically 5–15 residues long (Chiti et al. 2002a). Interestingly, in a recent study, Yagi-Utsumi et al. (2020) demonstrated the effect of gravity on amyloid fibril morphology and fibrillation kinetics of amyloid β through experiments under microgravity conditions. Further, mutations in APRs and their flanking residues alter protein aggregation propensity, aggregation kinetics, and also the morphologies of its aggregates (Sipe and Cohen 2000; López de la Paz and Serrano 2004). Insertion of such aggregating peptides in globular proteins triggers aggregation (Ventura et al. 2004). For example, Fig. 2 shows the Cryo-EM structure of an amyloid fibril formed by the lambda light chain (AL55) of the IGLV6-57 germline gene (Swuec et al. 2019). The fibrils were isolated from deposits of amyloid light-chain (AL) cardiac

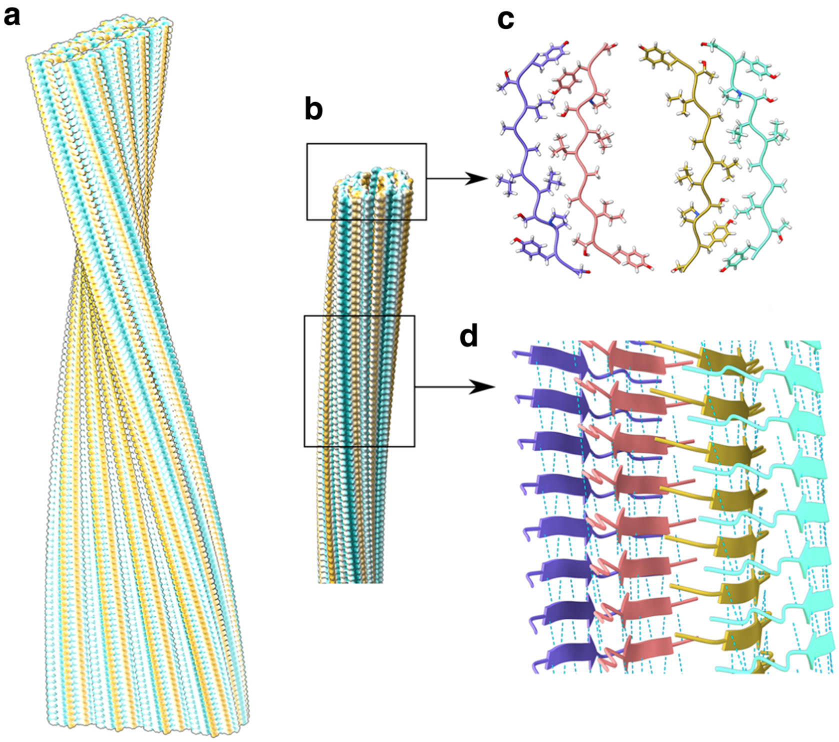

Fig. 1 Structural model of an amyloid fibril: a) amyloid fibril of an 11-residue fragment (125–135) of transthyretin protein (PDB: 3ZPK, UniProt ID: P02767), b) a protofibril , c) intersheet steric zipper formation, and d) the intersheet hydrogen bonds along the fibril axis

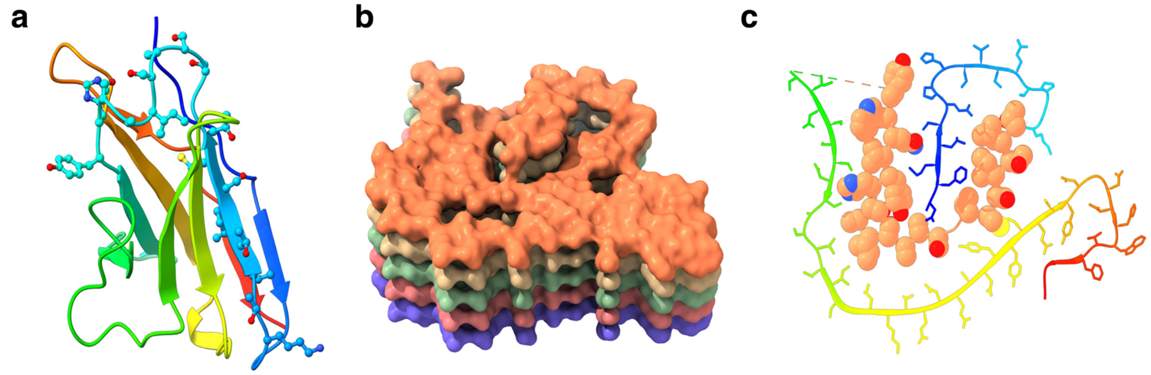

Fig. 2 Structure and predicted APRs in lambda light chain (A55): **a** structure of AL55 modelled using ABodyBuilder (Leem et al. 2016), **b** cryo-EM structure of AL55 amyloid protofibril (PDB: 6HUD), and **c** residue contacts in the fibril. The atoms constituting APRs (17–38) predicted using WALTZ, PASTA2, ANuPP, FishAmyloid, and MetAmyl (consensus) are highlighted as spheres. ChimeraX was used for visualization (Goddard et al. 2018)

amyloidosis patients through autopsy. The APRs identified by several APR prediction proteins cluster in the sequence region 17–38. The ability of the APRs to dictate the fates of even large proteins has attracted considerable research efforts.

Several computational methods have been developed to identify the aggregation-prone regions (APRs) in proteins and peptides and predict protein aggregation propensity (Meric et al. 2017) (Fig. 3). These methods can be classified based on their application and the features used to predict (Table 1). Broadly, these methods can be divided into sequence- and structure-based methods depending on the input data required for the prediction.

Sequence-based approaches to predict protein aggregation

Sequence-based approaches to predict aggregation in proteins rely on features such as amino acid physiochemical properties, sequence patterns, statistically derived propensity values, knowledge-based scoring functions, secondary structure propensities, residue-residue contact potentials, and threading. Pattern matching is the simplest of all the sequence approaches. Lopez de la paz and Serrano (2004) carried out experiments on peptide STVIIE through positional scanning mutagenesis and identified sequence patterns of hexapeptides, which form amyloid-like fibrils in vitro. However, the patterns

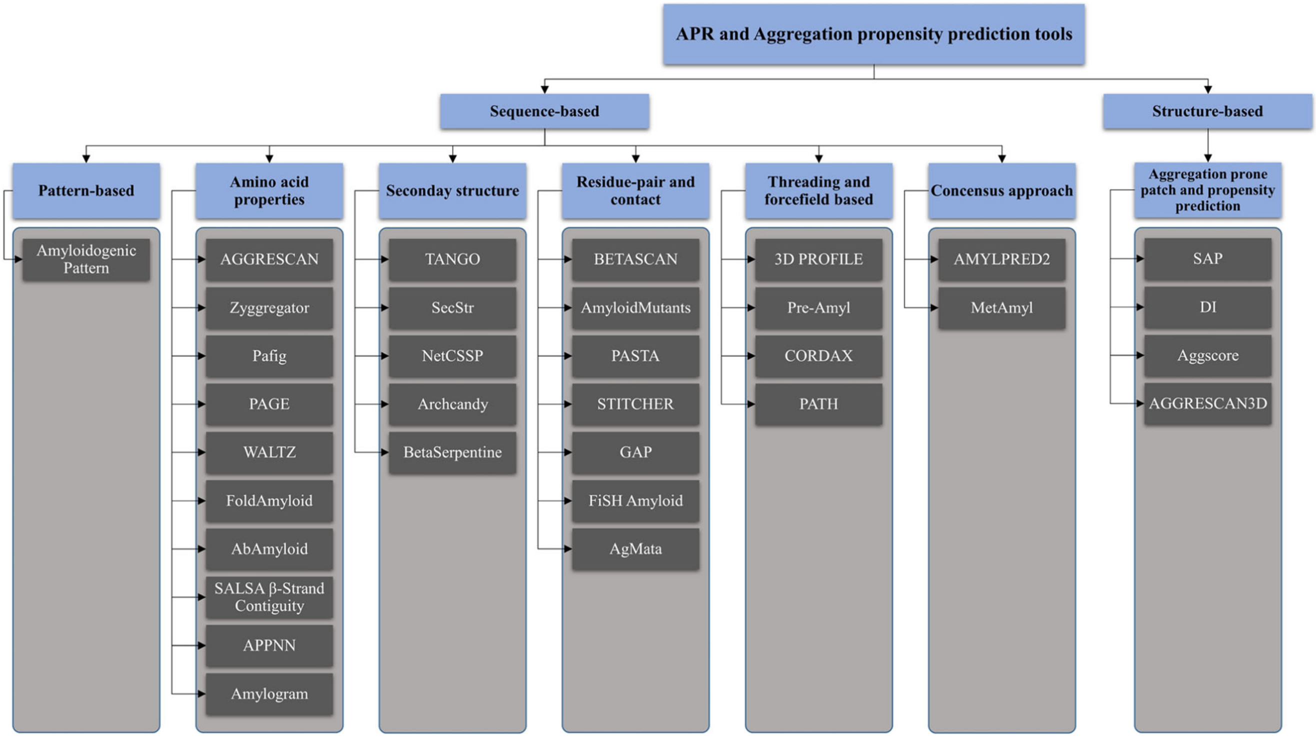

Fig. 3 Various APR and aggregation propensity prediction tools

Table 2 Summary of available methods to predict APR and aggregation propensity of proteins

Methods	Features	Application‡	Link	Reference
(a) Sequence-based features				
Sequence pattern				
Amyloidogenic pattern	Pattern derived from positional scanning mutagenesis experiments on amyloidogenic peptide STVIIE	APR	–	Lopez de la Paz and Serrano (2004)
Amino acid properties				
AGGRESCAN†	Aggregation propensity scale for amino acids derived from in vivo experiments on amyloidogenic proteins	APR, AP	http://bioinf.uab.es/aggrescan/	Conchillo-Solé et al. (2007)
Zyggregator	Amino acid scales for α-helix and β-sheet formation, hydrophobicity and charge, hydrophobic pattern, and presence of Gatekeeper residues	APR	–	Tartaglia and Vendruscolo (2008)
Pafig§,*	41 amino acid physicochemical properties	APR	http://www.mobioinfor.cn/pafig/	Tian et al. (2009)
PAGE	Aromaticity, β-propensity, charge, polar-nonpolar surfaces, and solubility	APR	–	Tartaglia et al. (2005)
WALTZ†,§	PSSM, physicochemical properties, position-specific pseudoenergy terms	APR	https://waltz.switchlab.org/	Maurer-Stroh et al. (2010)
AbAmyloid†,*	Amino acid composition, dipeptide composition, and physicochemical properties	AP	http://iclab.life.nctu.edu.tw/abamyloid	Liaw et al. (2013)
FoldAmyloid†	Packing density and hydrogen bond probability obtained from protein structures	APR, AP	http://bioinfo.protres.ru/fold-amyloid/	Garbuzynskiy et al. (2010)
SALSA β-Strand Contiguity (β-SC)	β-strand propensity	APR		Zibaee et al. (2007)
APPNN§	7 amino acid physicochemical and biochemical properties	APR	http://cran.r-project.org/web/packages/appnn/index.html	Família et al. (2015)
Amylogram†,§	17 amino acid properties such as size of residues, hydrophobicity, solvent surface area, frequency in β-sheets, contactivity, and contact site propensities	AP	http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/; http://github.com/michbur/AmyloGramAnalysis	Burdukiewicz et al. (2017)
ANuPP†	Atom compositions of peptides and protein segments	APR, AP	https://web.iitm.ac.in/bioinfo2/ANuPP/	Prabakaran et al. (2020)
Secondary structure propensity				
TANGO†,§	Segment β-sheet probability derived from empirical and statistically derived energy functions	APR, AP	http://tango.crg.es/	Fernandez-Escamilla et al. (2004)
SecStr†	Secondary structure preference	APR	http://biophysics.biol.uoa.gr/SecStr/	Hamodrakas et al. (2007)
NetCSSP†	Residue interaction and solvation energy obtained using AMBER forcefield	APR	http://cssp2.sookmyung.ac.kr/	Kim et al. (2009)
Archcandy§	Scoring function derived for steric tension, electrostatic interactions, packing, and hydrogen bond formation	Zipper	–	Ahmed et al. (2015)
BetaSerpentine†	β-arches (β-strand-loop-β-strand motif from Archcandy), compatibility of β-arches, compactness	Zipper	https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=25	Bondarev et al. (2018)
Residue-pair occurrence and contact preference				
BETASCAN†	Pairwise probability tables to identify hydrogen bond forming residues in strand pairs	APR	http://betascan.csail.mit.edu	Bryan et al. (2009)
AmyloidMutants†	Potential energy scoring function derived from observed residue/residue interactions in PDB	APR, Zipper	http://amyloid.csail.mit.edu/	O'Donnell et al. (2011)
STITCHER†,*	Scoring function addressing enthalpic and entropic changes in protofibril formation, and BETASCAN strand pair predictions	Zipper	http://stitcher.csail.mit.edu	Bryan et al. (2012)
PASTA 2†	Hydrogen-bonding energy function for residue pairs derived from beta-strand structures	APR, Zipper	http://old.protein.bio.unipd.it/pasta2/	Walsh et al. (2014)
GAP†		APR	http://www.iitm.ac.in/bioinfo/GAP/	

Table 2 (continued)

Methods	Features	Application‡	Link	Reference
	Residue pair potential derived from hexapeptide sequences			Thangakani et al. (2014)
FISH Amyloid†	Residue cooccurrence matrix derived from amyloidogenic and non-amyloidogenic peptides of length (4–10)	APR	http://comprec-lin.iiar.pwr.edu.pl/fishInput/	Gasior and Kotulska (2014)
AgMata§	Statistical potentials derived for residue position secondary structure probability and interaction energy	APR	https://bitbucket.org/bio2byte/agmata	Orlando et al. (2020)
Threading- and forcefield-based approach				
3D PROFILE (ZipperDB)†	Microcrystal structure of the NNQQNY peptide and atomic-level potential ROSETTADESIGN	Zipper, Zipper3D	https://services.mbi.ucla.edu/zipperdb/submit	Thompson et al. (2006)
Pre-Amyl*	Template ensemble obtained from microcrystal structure of the NNQQNY peptide and KBP, atom distance-dependent knowledge-based residue pairwise potential	Zipper, Zipper3D	ftp://mdl.ipc.pku.edu.cn/pub/software/pre-amyl/	Zhang et al. (2007)
CORDAX†	Thermodynamic stability calculated by threading over 140 amyloid fibril cores	Zipper, Zipper3D	https://cordax.switchlab.org/	Louros et al. (2020)
PATH	Modeller Dope score and Rosetta (REF15) energy values from homology models of 7 template structures	Zipper, Zipper3D	https://github.com/KubaWojciechowski/PATH	Wojciechowski and Kotulska (2020)
Consensus approach				
AMYLPRED2†	Consensus predictor includes outputs from AGGRESCAN, NetCSSP, AmyloidMutants, Pafig, Amyloidogenic Pattern, SecStr, Average Packing Density , TANGO, Beta-strand contiguity, WALTZ, Hexapeptide Conformational Energy	APR	http://aias.biol.uoa.gr/AMYLPRED2/	Tsolis et al. (2013)
MetAmyl†	Consensus predictor includes PAFIG, SALSA, WALTZ, and FoldAmyloid	APR	http://metamyl.genouest.org/	Emily et al. (2013)
(b) Structure-based features				
Accessible surface area and surface patches				
SAP	Residue hydrophobicity, solvent accessible area over time obtained from MD	APP	–	Chennamsetty et al. (2009)
Developability index	SAP and PROPKA values	AP	–	Lauer et al. (2012)
Aggscore	Hydrophobic and hydrophilic patches obtained using atom partial charges and logP values	APR	–	Sankar et al. (2018)
AGGRESCAN3D 2.0†,§	AGGRESCAN residue score, exposed surface area, FoldX energy-minimized protein structure or Ensemble from CABS-flex simulations	AP	http://biocomp.chem.uw.edu.pl/A3D2/; https://bitbucket.org/lcbio/aggrescan3d	Kuriata et al. (2019)

† Available as webserver

§ Available as stand-alone tool

* Server or stand-alone is not available/accessible (last checked: Sep 2020)

‡ Applications were grouped into (a) APR: identification of APR and peptides in protein sequence, (b) Zipper: prediction of Zipper–strand orientation and residue pairing, (c) Zipper 3D: prediction or modelling of protofibril or cross-β spine models, and (d) AP: quantification of protein aggregation propensity

are specific to certain hexapeptides and did not cover several new amyloid-forming peptides that were reported over the years.

Amino acid properties Amino acid properties such as hydrophobicity, size, surface area, charge, aromaticity, contact frequency, beta-sheet propensity, and several other physicochemical properties are used for the identification of aggregation-prone regions. AGGRESCAN uses the amino acid aggregation-propensity scale derived from in vivo experiments on amyloidogenic proteins (Conchillo-Solé et al. 2007). These propensity values were used to identify the aggregation-prone regions through a sliding-window approach. Zyggregator uses amino acid scales for α-helix and β-sheet formation, charge, hydrophobicity, and also hydrophobic pattern and presence of gatekeeper residues (Tartaglia and Vendruscolo 2008). WALTZ uses a hybrid approach, combining a position-specific scoring matrix derived from amyloidogenic peptides and amino acid physicochemical properties and position-specific pseudoenergy values obtained from modelled structures (Maurer-Stroh et al. 2010). ANuPP is an ensemble classifier that consists of nine logistic regression models trained independently on groups of amyloidogenic peptides to address the diversity in aggregation nucleation, propagation, and fibrillation processes. ANuPP uses atom composition as features to represent sequence segments (Prabakaran et al. 2020).

Secondary structure preference The propensity to form β-sheet is one of the key features of amyloid-fibril-forming peptides and proteins, and this has been extensively used in developing prediction algorithms. TANGO uses various empirically and statistically derived potential functions to estimate the probability of a segment to form β-strand-mediated aggregates (Fernandez-Escamilla et al. 2004). In theory, TANGO compares the probability of a segment to be in various secondary structural states such as α-helix, β-sheet, coil, and turn. Conformational switch from other secondary states to β-sheet formation is the principle behind SecStr and NetCSSP (Hamodrakas et al. 2007; Kim et al. 2009).

Residue-pair occurrence and contact preference Cross-β spine made of steric zippers is a common feature of all amyloid fibrils. These steric zippers consist of interdigitating side chains and axial hydrogen bonds and strengthen the supramolecular structure of amyloid fibrils. In addition, the stacking of aromatic residues and ladders of hydrogen bonds formed by Asn, Gln, Thr, and Ser residues adds additional stability to the structure. These residue-residue interactions are seen as crucial for an APR and thus used in various prediction methods. GAP uses position-specific residue-pair energy potential derived from amyloid and amorphous β-aggregating hexapeptide sequences to identify amyloidogenic peptides (Thangakani et al. 2014). PASTA2 and BETASCAN use residue-residue probabilities and scoring functions for β-sheet hydrogen bond formation and contact derived from protein structure databases. Apart from predicting the APR stretch on the protein sequence, these approaches can predict the β-strand orientation and pairing between residues.

Threading and forcefield Thompson et al. (2006) used the crystal structure of the cross-β spine of peptide NNQQNY to identify APR segments in amyloidogenic proteins. Each hexapeptide from a given protein sequence was mapped onto an ensemble of steric zipper templates and scored subsequently. The assumption behind this 3D profile method is the conserved cross-β motif in amyloid fibrils of diverse proteins. Apart from identifying amyloidogenic peptides and regions in protein sequence, the approach is capable of predicting orientation between strands forming the zipper.

Structure-based approaches

Structure-based methods such as SAP, developability index, AGGRESCAN3D, and Aggscore require protein structure as input (Chennamsetty et al. 2009; Lauer et al. 2012; Zambrano et al. 2015; Sankar et al. 2018). These methods predominantly take account of the solvent accessibility of protein residues and atoms to estimate surface hydrophobicity. In addition to the static structure, short molecular dynamics (MD) simulations are performed to calculate the ensemble statistic over time. Unlike sequence-based methods, structure-based methods account for the folding and native state of a protein. At the same time, the limited timescale of MD simulations also biases the prediction of a single protein structure in its native state. This approach may not hold for highly dynamic proteins with multiple metastable states and disorder regions.

Comparison of APR prediction tools

Table 2 lists the various APR and aggregation propensity prediction tools that have been published to date. We selected ten prediction tools from the literature that were accessible as a webserver or stand-alone application and easily applicable to large datasets for a comparative evaluation. Table 3 lists the performance of these APR prediction tools in distinguishing between APR and non-APR segments in a dataset of 37 amyloidogenic proteins. The dataset was extracted from AmyPro database (Varadi et al. 2018) at 40% sequence identity cut-off. Segment OVerlap (SOV) scores: $\mathrm{SOV}_{\mathrm{APR}}$, $\mathrm{SOV}_{\text{non-APR}}$, $\mathrm{SOV}_{\mathrm{average}}$, and $\mathrm{SOV}_{\mathrm{overall}}$ are used for evaluation (Zemla et al. 1999). Similar to secondary structure assessments, SOV scores the prediction performance based on the overlap between the predicted and actual segments instead of residue-wise comparison. Overall, the consensus methods, Amylpred2 and MetAmyl, showed better performance over other methods. ANuPP and TANGO scored better than other methods with $\mathrm{SOV}_{\mathrm{overall}}$ of 50.2 and 48.1, respectively. Though several tools showed a good overall score, they exhibited an imbalance between $\mathrm{SOV}_{\mathrm{APR}}$ and $\mathrm{SOV}_{\text{non-APR}}$. Similar assessments were performed based on a dataset of 142 amyloid-like fibril-forming hexapeptides from WALTZ-DB 2.0 (Louros et al. 2020). In spite of the imbalance in sensitivity and specificity, ANuPP and AGGRESCAN scored better than other methods (data not included). These results highlight the need for more robust methods to identify APRs accurately.

Table 3 Performance of APR identification algorithms and tools

	SOV_{APR}	$SOV_{non-APR}$	$SOV_{overall}$	$SOV_{average}$
AGGRESCAN	34.3	36.5	32.4	35.4
Amyloidogenic pattern	14.9	53.4	44.1	34.2
Amylpred2	29.5	41.1	34.8	35.3
ANuPP	45.2	52.3	50.2	48.7
FishAmyloid	14.5	45.2	37.5	29.9
MetAmyl	43.1	33.4	31.5	38.2
NetCSSP	32.9	35.9	32.7	34.4
Pre-Amyl	32.9	36.4	34.3	34.7
FoldAmyloid	30.2	41.5	35.2	35.8
Pafig	30.9	25.8	25.2	28.3
Pasta2 (85% specificity)	13.2	24.9	23.2	19.1
Pasta2 (90% specificity)	12.2	25.7	23.6	18.9
SALSA Strand contiguity	29.8	45.1	39.4	37.5
SecStr	12.0	47.6	38.5	29.8
TANGO	19.1	57.8	48.1	38.5
WALTZ	44.4	28.9	28.7	36.6

Segment OVerlap scores (SOV_{APR}, $SOV_{non-APR}$, $SOV_{overall}$, and $SOV_{average}$) are used to evaluate the prediction of APRs in proteins (Zemla et al. 1999). SOVaverage was calculated as an average of SOV_{APR} and $SOV_{non-APR}$

Aggregation kinetic prediction tool

Aggregation kinetics measure how fast/slow the proteins will aggregate under given experimental conditions. The aggregation mechanisms, curve fitting, and experiments related to aggregation kinetics have been reviewed earlier in the literature (Morris et al. 2009; Hirota et al. 2019). In this section, we have focused on the role of biophysical features and experimental conditions in determining aggregation kinetics.

The detailed in vitro analysis of mutants of acylphosphatase (AcP) revealed the role of charge, hydrophobicity, and secondary structure propensity towards altering aggregation kinetics (Chiti et al. 2002a; Chiti et al. 2002b). Further, they derived the first empirical equation to predict the change in aggregation kinetics upon point mutation using the physicochemical features of proteins (Chiti et al. 2003). Subsequently, several studies on different amyloidogenic proteins analysed the role of physicochemical properties on protein aggregation such as hydrophobicity (Calamai et al. 2003; Fink 1998; Hilbich et al. 1992), β-strand propensity (Tartaglia et al. 2004; Família et al. 2015; Tjernberg et al. 2002; Fernandez-Escamilla et al. 2004), polarity (Tartaglia et al. 2004; Polanco et al. 2015), charge (Tartaglia et al. 2004; Calamai et al. 2003; Tjernberg et al. 2002), aromaticity (Tartaglia et al. 2004; Azriel and Gazit 2001; Gazit 2002), and stability (Fink 1998; Ramírez-Alvarado et al. 2000; Brito et al. 2003).

The aggregation kinetics assays have shown that the rate of aggregation is sensitive to even a small change in experimental conditions such as protein or buffer concentration, pH, temperature, ionic concentration, seeding, or agitation (Brudar and Hribar-Lee 2019; Hortschansky et al. 2005; Morel et al. 2010; Ow and Dunstan 2013). Currently, there are few in silico methods available to predict the absolute aggregation rate or change in aggregation rate upon point mutation (Table 4) as discussed below.

Methods to predict change in aggregation rate upon point mutation

Chiti et al. (2003) first proposed a mathematical equation (Eq. 1) to predict the change in aggregation rate using intrinsic protein sequence features, which includes change in the hydrophobicity of the polypeptide chain (ΔHydr.), propensity to convert from α-helical to β-sheet structure ($\Delta\Delta G_{coil-\alpha}+\Delta\Delta G_{\beta-coil}$), and change in overall charge (ΔCharge).

$$\ln\left(\frac{\upsilon_{mut}}{\upsilon_{wt}}\right) = A\Delta\text{Hydr.} + B\left(\Delta\Delta G_{coil-\alpha} + \Delta\Delta G_{\beta-coil}\right) + C\Delta\text{Charge} \qquad (1)$$

where *A*, *B*, and *C* in the above equation are constants, which are estimated by fitting the equation to experimental change in the aggregation rate. The model achieved a correlation of 0.85 on a set of 27 mutations found in short peptides or natively unfolded proteins, including amylin, amyloid β-peptide, tau, and α-synuclein. This model has some limitations, including (i) smaller dataset size, (ii) inability to predict aggregation

Table 4 Summary of available methods to predict aggregation kinetics

Prediction model	Prediction for	Performance (training dataset)*	Availability
Chiti's (Chiti et al. 2003)	Point mutation	r=0.85 (27)	Eq. 1
DuBay's (DuBay et al. 2004)	Protein/peptides	r=0.92 (79)	Eq. 2
Tartaglia's (Tartaglia et al. 2005)	Protein/peptides	r=0.95 (90)	Eqs. 3–6
Yang's (Yang et al. 2019)	Protein/peptides	r ~0.96 (140)	–
AggreRATE-Disc (Rawat et al. 2018)	Point mutation	84% (220)#	https://www.iitm.ac.in/bioinfo/aggrerate-disc/
AggreRATE-Pred (Rawat et al. 2020b)	Point mutation	r ~0.82 (183)	https://www.iitm.ac.in/bioinfo/aggrerate-pred/
AbsoluRATE (personal communication)	Protein/peptides	r=0.74 (82)	https://web.iitm.ac.in/bioinfo2/absolurate-pred/

*Performance of the model described by the respective model for their respective training dataset. The performance measure r denotes correlation

# Percentage of correctly predicted aggregation rate enhancer or mitigator mutations (accuracy)

kinetics for mutations involving proline residues due to undefined values for change in β-sheet propensity ($\Delta\Delta G_{\beta-\text{coil}}$), and (iii) inability to predict the aggregation kinetics for residues, for which α-helical propensity is predicted zero by the AGADIR server (Muñoz and Serrano 1994).

Rawat et al. (2018) developed the method AggreRATE-Disc (Discrimination of Aggregation Rate change Upon Mutation) using sequence-based features to predict the aggregation rate enhancer or mitigator mutations using machine learning. It is developed using a support vector machine (SVM)–based classifier on 220 point mutations from 25 proteins. The model grouped the mutations based on the local secondary structure conformation at the mutation site (helix, strand, and coil) and achieved an average prediction accuracy of ~82% using leave-one-out cross-validation. AggreRATE-Disc identified a unique set of sequence-based features that influence the aggregation rate in each mutation site conformational class. For example, changes in protein stability and flexibility in the helical region influence the rate of aggregation. Similarly, the aggregation rate is mainly affected by charge, polarity, and β-strand propensity when the mutations fall in the β-strand regions. For other mutation sites falling under the coil category, such as bends, turns, and disordered regions, aggregation rates are affected by both helical tendency and aggregation propensity.

The AggreRATE-Pred model (Rawat et al. 2020b) was an improvement over AggreRATE-Disc, which included structure-based features to predict the quantitative change in aggregation rate. The statistical model is developed by combining four different regression equations, which is generated by classifying the data based on polypeptide length and local secondary structure conformation at the mutation site, and fitting of the regression equation. The dataset of 183 point mutations in 23 amyloidogenic proteins was primarily divided into two groups: (i) short peptides (length < 40 residues) and (ii) long polypeptides and proteins (length ≥ 40 residues). The long polypeptide and protein dataset are further classified to helix, strand, and coil class based on local secondary structure conformation, similar to the previous study (Rawat et al. 2018). The statistical model achieved an average correlation coefficient of ~0.82 and an average MAE of ~0.43 on the training dataset. The regression analysis showed the importance of local structural context, thermodynamic stability changes, and effect of neighbour residues at the mutation site.

Methods to predict the absolute aggregation rate

DuBay et al. (2004) improved Eq. 1 to predict the absolute aggregation rate of polypeptides using intrinsic features, such as hydrophobicity (I^{hydr}), alternating hydrophobic-hydrophilic residue pattern (I^{pat}), and absolute value of net charge (I^{ch}), and extrinsic features, pH (E^{pH}), ionic strength (E^{ionic}), and polypeptide concentration (E^{conc}). The mathematical formula for the prediction of the absolute rate of aggregation is given in Eq. 2.

$$\log(k) = \alpha_0 + \alpha_{\text{hydr}} I^{\text{hydr}} + \alpha_{\text{pat}} I^{\text{pat}} + \alpha_{\text{ch}} I^{\text{ch}} + \alpha_{\text{pH}} E^{\text{pH}} + \alpha_{\text{ionic}} E^{\text{ionic}} + \alpha_{\text{conc}} E^{\text{conc}} \quad (2)$$

where log(k) is the logarithm in base 10 of the aggregation rate (k) in units of s^{-1}. α values are constants estimated by fitting the equation on experimental data of 79 mutations. The model has achieved a correlation of 0.92 on the training dataset of 79 proteins/peptides. However, the model was prone to biasness due to limited availability of the aggregation rates, where 59 out of 79 data points were point mutation variants of acylphosphatase (AcP) protein.

Tartaglia et al. (2005) proposed a sequence-based algorithm to predict the aggregation rate and aggregation-prone regions in protein/polypeptide sequences. The aggregation propensity (π_{il}) of the sequence is calculated using position-dependent factors (Φ_{il}) and composition-dependent factors (φ_{il}).

$$\pi_{il} = \Phi_{il} \varphi_{il} \quad (3)$$

where l is the length of the segment starting at the position i in the sequence. The position-dependent factors include aromaticity (A_{il}), β-propensity (B_{il}), and charge (C_{il}).

$$\Phi_{il} = e^{A_{il}+B_{il}+C_{il}} \tag{4}$$

The amino acid composition–dependent factors include side-chain-accessible surface area of apolar (S_j^{a}), polar (S_j^{p}), and all residues (S_j^{t}); solubility (σ_j); and parallel (⇈) and antiparallel (⇅) tendency to aggregate. The hatted values are averages of 20 standard amino acids.

$$\varphi_{il} = \left[\prod_{j=1}^{i+l-1}\left(\frac{S_j^{\mathrm{a}}}{\widehat{\mathrm{S}}^{\mathrm{a}}}\theta\uparrow\uparrow + \frac{S_j^{\mathrm{p}}}{\widehat{\mathrm{S}}^{\mathrm{p}}}\theta\uparrow\downarrow\right)\frac{\widehat{\mathrm{S}}^{\mathrm{t}}}{S_j^{\mathrm{t}}}\,\frac{\widehat{\sigma}}{\sigma_j}\right]^{1/l} \tag{5}$$

The model predicts the aggregation rate from the aggregation propensity (Eq. 3) by including a function for experimental conditions ($\alpha(c, T)$, which takes account of protein concentration and temperature) as given in the following formula (Eq. 6):

$$v_{il} = \alpha(c, T)\pi_{il} \tag{6}$$

The model achieved a correlation of 0.95 with 90 data points. However, this model also suffers from the limited availability of the data and biasness within the dataset.

Yang et al. (2019) proposed a feedforward fully connected neural network (FCN)–based machine learning model for predicting the absolute aggregation rates. The model is trained on a dataset of 21 amyloidogenic proteins (140 data points) using 16 intrinsic sequence-based features and 4 extrinsic features. The model focuses on the inclusion of more experimental conditions and considers them as a separate data point in the prediction model. Although the model showed an average prediction accuracy of more than 90% on the training dataset, it seems overfitted as it employs 16 intrinsic sequence-based features to essentially predict 21 sequence variants of amyloidogenic proteins.

AbsoluRATE (Rawat et al., manuscript under review) is a support vector machine (SVM)–based regression model to predict absolute rates of protein and peptide aggregation. The model trained on 82 non-redundant proteins/peptides has achieved a correlation coefficient of 0.72 with MAE of 0.91 (natural log of k_{app}, where k_{app} is in hour^{-1}) using leave-one-out cross-validation. The model accounts for sequence-based features (such as features derived from APR prediction servers, disorderness, polarity, beta-sheet propensity, etc.) and extrinsic features (such as temperature, pH, ionic and protein concentration).

Comparison of aggregation kinetics prediction methods

The limited availability of experimental aggregation kinetics resources is a major limitation towards the development of accurate computational models. Hence, to benchmark the above kinetic models, we randomly selected a test set from the AggreRATE-Pred dataset (Rawat et al. 2020b) in such a way that it (i) includes all structural classes (10% of the training dataset in the respective class) and (ii) has predictable aggregation rates for all the participating models (Table 5). For a fair comparison, we retrained the AggreRATE-Pred by removing test set data points and achieved a correlation of 0.81 on the training dataset (original correlation $r =$ 0.82). The performance of the removed test set was further evaluated on the newly developed AggreRATE-Pred model with a reduced training set, as shown in Table 5. The absolute aggregation rate prediction models were tested by subtracting the predicted aggregation rates for mutant and wild-type protein sequences. The correlations obtained by the absolute aggregation rate prediction models were expectedly low, with the highest correlation of 0.41 obtained by Tartaglia's model (Tartaglia et al. 2005). Chiti's and DuBay's models are almost two-decade-old models trained on minimal datasets available at that time, which is also reflected in the prediction performance of these methods on the test dataset. AggreRATE-Pred, a structure-based method, showed the highest correlation among all models. AggreRATE-Disc is a sequence-based method that cannot predict the quantitative change in aggregation rate. However, it has correctly predicted the effect on aggregation rate (increase/decrease) with 73.7% accuracy. Yang's model (Yang et al. 2019) was not benchmarked due to the unavailability of the webserver/stand-alone program.

Molecular dynamics approach

A better understanding of the physical phenomenon and mechanistic details of self-association is often obtained through molecular simulation studies. Molecular dynamics and Monte Carlo simulations have been widely used to understand protein aggregation dynamics to study various aspects. Various simulation techniques and methodologies have been used to study protein and peptide aggregation. Depending on the focus of the study, the simulation can vary from (i) coarse-grained to all-atom models, (ii) Monte Carlo to molecular dynamic simulations, (iii) implicit to explicit solvation models, and (iv) dimers to bulk simulations (Morriss-Andrews and Shea 2014, 2015; Carballo-Pacheco and Strodel 2016). APRs from the proteins are often studied as peptides instead of the whole protein to understand the aggregation mechanisms and residue-residue interactions at a reduced computational cost. We have provided a summary of the diverse simulations and their applications below.

Table 5 Different kinetics prediction methods benchmarked on the test dataset

Protein information				Change in aggregation rates as $\Delta\ln(k_{app})$ predicted by						
Protein	Length	Mutation	Structural class	Experimental	Prediction model					
					Chiti	DuBay	Tartaglia	AggreRATE-Disc	AggreRATE-Pred	AbsoluRATE
AChE	14	S 8 A	Short peptide	− 1.66	− 1.04	0.64	− 0.17	Decrease	− 1.39	0.19
AChE	14	H 4 A	Short peptide	− 0.83	− 0.59	0.05	1.28	Decrease	− 0.96	0.38
AChE	14	F 3 A	Short peptide	− 1.56	2.03	0.21	− 3.09	Decrease	− 1.41	− 0.36
α-Synuclein	140	H 50 A	Helix	− 0.12	− 0.39	− 0.17	− 0.6	Decrease	0.18	0.02
Aβ40	40	A 21 C	Helix	− 0.01	− 0.51	0.01	0.15	Decrease	− 0.59	1.05
Acylphosphatase	98	E 29 R	Helix	− 1.54	0.91	− 0.3	− 1.47	Increase	− 1.16	0.05
Barstar	89	S 14 C	Helix	0.63	− 1.1	− 0.03	0.05	Increase	− 0.35	1.18
Acylphosphatase	98	F 94 L	Strand	− 0.09	1.47	0.01	− 0.04	Decrease	0.56	− 0.04
Acylphosphatase	98	V 9 A	Strand	0.13	1.49	0.01	− 0.04	Decrease	− 0.05	0.08
AL-12	108	R 65 S	Strand	0.92	− 1.36	0.12	0.5	Increase	1.02	− 0.09
Stefin B	98	G 50 E	Strand	− 0.69	0.66	0.21	0.68	Decrease	− 0.48	0.06
Aβ40	40	E 3 R	Coil	0.08	0.66	− 0.28	− 1.73	Increase	− 0.41	0.09
Aβ42	42	D 23 N	Coil	− 0.33	− 1.61	− 0.23	− 0.55	Increase	− 0.44	0.98
Aβ42	42	E 11 K	Coil	− 0.35	− 0.15	− 0.33	− 1.68	Decrease	− 0.49	0.01
Aβ42	42	D 7 N	Coil	− 0.48	− 1.55	− 0.23	− 0.55	Decrease	0.14	− 0.02
Aβ42	42	E 22 G	Coil	− 1.51	− 0.68	− 0.27	− 0.72	Increase	− 0.61	− 0.05
Acylphosphatase	98	L 65 V	Coil	0.02	− 0.18	0.01	− 0.03	Increase	− 0.38	0
Acylphosphatase	98	G 45 E	Coil	1.09	0.67	0.21	0.63	Decrease	0.67	0.06
Insulin	51	T 8 H	Coil	− 0.77	0.72	0.13	0.16	Decrease	− 0.2	− 1.75
Performance[#]					*-0.14*	*-0.08*	*0.41*	*73.7%*	*0.80*	*0.26*

The test dataset used in the study is taken from the AggreRATE-Pred server

[#] The correlation between experimental and predicted change in aggregation rates from the respective server (percentage of correct prediction as increase/decrease in aggregation rate upon mutation for AggreRATE-Disc)

Atomic simulation of peptide assembly

All atom simulations are often limited to the simulation of peptide aggregation. For example, Ma and Nussinov (2002a) carried out molecular dynamics simulations of the two peptides, AGAAAAGA observed in PrP protein and a polyalanine peptide AAAAAAAA to identify critical oligomer size (Ma and Nussinov 2002a). They showed that oligomers of size 6–8 strands were found to be stable and retained the fibril model conformation (10 Å inter-sheet distance and 5 Å inter-strand distance). Similar works have been carried out on several peptides such as poly-glutamine, poly-glycine, and peptide fragments from amyloidogenic proteins (Cecchini et al. 2006; Karandur et al. 2014; Marchut and Hall 2006). Ma and Nussinov (2002b) also studied three different segments of amyloid β (16–22, 6–35, and 10–35) using MD and compared the results with solid-NMR structures. Gsponer et al. (2003) studied the heptapeptide GNNQQNY from Sup35 using 20-ns simulations of a 3-peptide system and observed 25 parallel β-strand formation events, consistent with experimental data. The study showed the influence of residues on orientation preference and stability of the strand. Zanuy et al. (2003) studied the oligomeric stability of two segments (NFGAIL 22–27 and NFGAILSS 22–29) from islet amyloid polypeptide using molecular dynamics. Their work highlighted the importance of the assembly of interacting sheets in amyloid fibril formation (Zanuy et al. 2003; Zanuy and Nussinov 2003). Similar studies have been carried out on peptides STVIIE and its 5 variants, NFGAIL 22–27 of the human islet amyloid polypeptide, and hIAPP 1–19 peptide (Wu et al. 2005; López De La Paz et al. 2005; Guo et al. 2015; Tran and Ha-Duong 2015). Priya and Gromiha (2019) revealed that the length of polyQ in the aggregation of huntingtin protein is important for β-sheet formation and for elucidating the pathological mechanism in Huntington disease. Figure 4 shows the first and last snapshots from the multi-copy MD simulations of the ‘VLVIY’ peptide assembly. Studies showed that introducing the lysine residue in the ‘VLVIY’ segment increased the solubility and

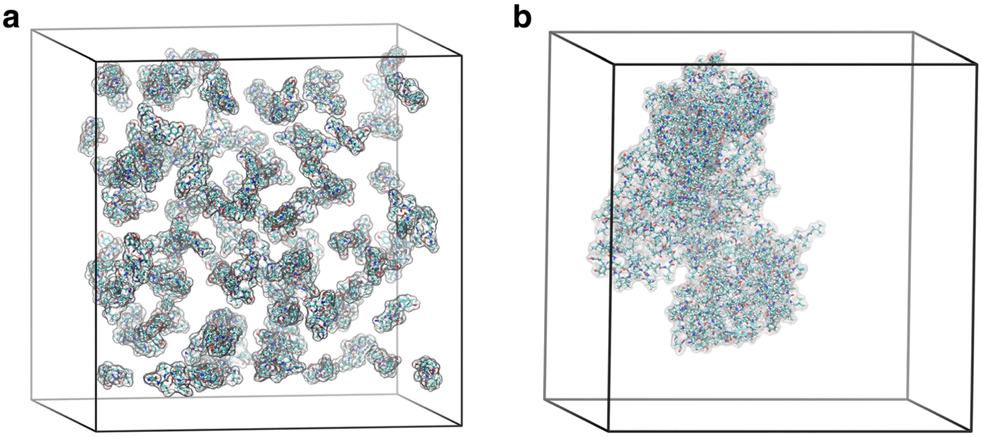

Fig. 4 All-atom simulation of peptide (VLVIY) assembly: **a** the initial setup of 105 peptides separated by 1 nm distance from each other and **b** the aggregated peptides after a simulation time of 50 ns. VMD was used for the visualization (Humphrey et al. 1996)

reduced the viscosity of a monoclonal antibody, stamulumab (Nichols et al. 2015; Kumar et al. 2018).

Extending the spatio-temporal limits using coarse-grained models

Coarse-grained models are simplified models of polypeptides and their associated interactions. Depending on the depth of abstraction and the level of resolution, coarse-grained models can be categorized as phenomenological models, lower-resolution representative models, and high-resolution coarse-grained models (Morriss-Andrews and Shea 2015). The loss in detail and accuracy of a model is compensated with increased computational efficiency and inference from the extended spatio-temporal scale of simulations. Coarse-grained models extend the timescale of simulation beyond what is currently possible for the atomistic simulations, thereby assisting in studies of protein aggregation mechanisms, phase separation, and nanostructure formation. Such simulations can also help derive thermodynamic parameters of phase separation and fibril growth. All-atom and coarse-grained models of α-synuclein protein are shown in Fig. 5. Advantages of coarse-grained models to extend the spatio-temporal limitations of MD were put to use by Nguyen et al. (Nguyen and Hall 2004a, b, 2005, 2006). Nguyen and Hall (2004a) studied the phase diagram of the polyalanine peptide system using DMD simulations. The authors constructed five 96-peptide simulations at various concentrations using the PRIME model of the 16-residue polyalanine peptide. They showed that the peptide exists in four distinctive single-phase regions: α-helices, fibrils, nonfibrillar β-sheets, and random coils depending on concentration and temperature.

Marchut and Hall (2006, 2007) studied the aggregation of polyglutamine and the role of side chains using an intermediate resolution model, PRIME, which showed the spontaneous formation of long annular tube-like structures. Peng et al. (2004) studied the stacking of the entire amyloid β 1–40 peptide into β-sheet using discrete molecular simulation (DMD) and coarse-grained modelling and showed that the peptide system formed the stacking of β-strands at higher temperatures and amorphous aggregates at lower temperatures. Bellesia and Shea (2007, 2009) performed off-lattice simulations of peptide aggregation using a coarse-grained model, consisting of 2 and 1 beads representing backbone and C^{β}, respectively, and analysed the kinetics, thermodynamics, and aggregate structure through simulations of different peptide sequences. Interestingly, their work highlighted the role of charged residues in stabilizing and changing the preference of orientation of peptides during aggregation.

Singh et al. (2008) studied the effect of finite system size of peptide aggregation by simulating all an atom-model of the IAPP fragment (15–19) in the TIP3P water model using the AMBER force field and discussed the effect of concentration and system size on peptide aggregation. Magno et al. (2010) studied the effect of molecular crowding on the aggregation of an amphipathic peptide model through simulation of a 125-peptide system of varying box size (150 to 290 Å). They reported that crowders play a crucial role in accelerating the nucleation of low-aggregation propensity peptides. Matthes et al. (2011, 2012) studied the spontaneous steric zipper oligomerization of peptides 306–311 of tau protein, 12–17 of insulin B chain, and 51–56 segment of alpha-synuclein using an all-atom model of a 10-peptide system. Kumar et al. (2019) used MD simulation to analyse the aggregation propensity of the three peptides in 24–33 (N-terminal domain), 126–136 (RNA recognition motif 1), and 247–254 (RNA recognition motif 2) of human TDP-43. Wang et al. (2019) studied the solubility of different oligomers and fibril models of amyloid-beta (16–22) by measuring the equilibrium monomer concentration in the system using the PRIME20 model.

Molecular dynamics simulations have also been used to study the solubility and aggregation propensity of peptides (Karandur et al. 2014). Frederix et al. (2011, 2015) explored the self-assembly of the entire sequence space of dipeptides (400) and tripeptides (8000) through coarse-grain simulations using the MARTINI force field to measure the aggregation propensity of the peptide and the nature of nanostructures.

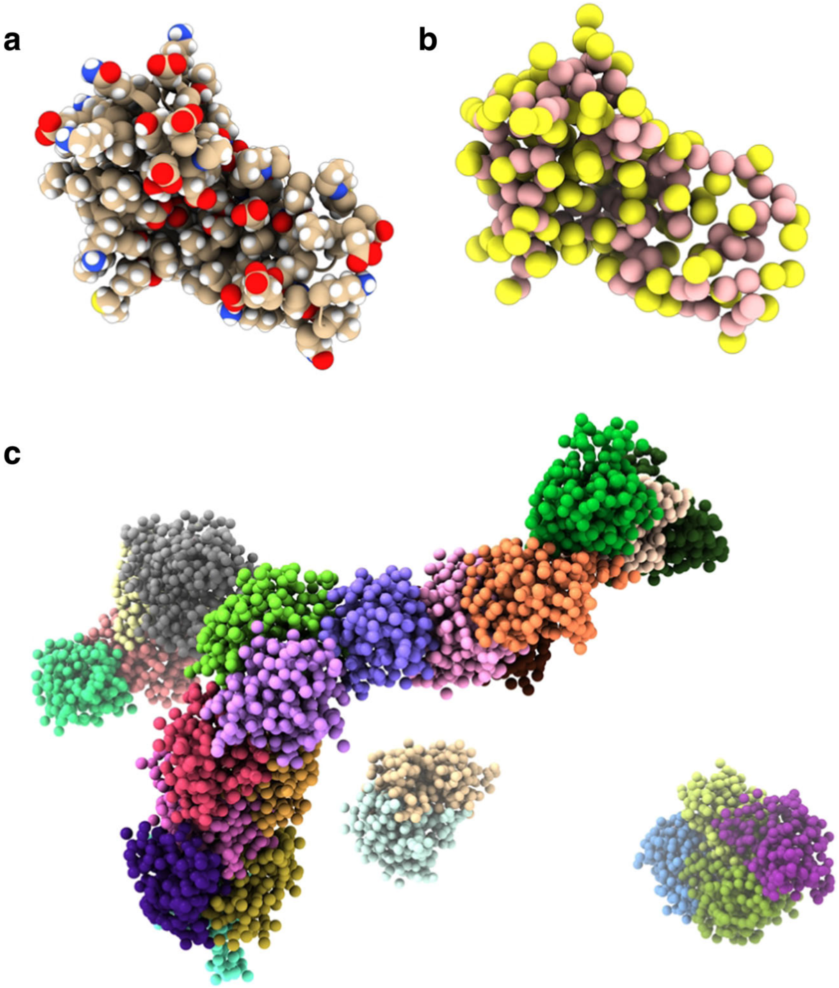

Fig. 5 Coarse-grained simulation of protein aggregation: **a** all-atom model, **b** Martini coarse-grained model (Marrink et al., 2007), and **c** aggregated structure obtained from coarse-grained simulation of α-synuclein protein. ChimeraX was used for visualization (Goddard et al. 2018)

They showed that aggregation propensity depends on hydrophobicity.

Understanding protein-protein interaction and oligomerization

Simulations of protein-protein interactions and protein oligomerization provide valuable insights on the role of residue-residue interactions, key structural motifs, and transitions in dictating protein aggregation at an earlier stage. Brown and Bevan (2016) investigated the oligomerization of amyloid-β and its binding to membrane models through simulation of a united-atom model with tetramer and pentamer systems. They explored the structural changes during oligomer-membrane binding to understand the Aβ oligomer toxicity. Similar studies using MD have been employed to study the formation and stability of oligomers of aggregation-prone protein such as amyloid-β, TDP-43, and amylin and the role residue-residue contact in oligomerization (Kumar et al. 2019; Berhanu and Masunov 2014; Khatua and Bandyopadhyay 2017). However, sampling the entire landscape of protein-protein aggregation in an explicit solvent model is computationally intensive on both by the spatial and temporal scales. Alternatively, coarse-grained models, implicit solvent models, and peptide simulation have been widely used in the literature (Morriss-Andrews and Shea 2014; Carballo-Pacheco and Strodel 2016). Molecular dynamic simulations were also carried out on immunoglobins to study the self-aggregation tendency of the molecules (Buck et al. 2013, 2015; Tiller et al. 2017).

Beyond coarse-grained models, continuum modelling has also been used to extend the spatio-temporal scale. Continuum models have also been developed to study the mesoscale properties of fibrils (Knowles and Buehler 2011; Paparcone et al. 2011). In addition, several techniques such as replica exchange molecular dynamics, Hamiltonian replica-permutation molecular dynamics, umbrella sampling, and metadynamics have also been applied to tune, accelerate, and study protein aggregation simulations (Barducci et al. 2006; Larini and Shea 2012; Itoh and Okumura 2013, 2016; Zheng et al. 2016; Morriss-Andrews and Shea 2014, 2015; Carballo-Pacheco and Strodel 2016). The Markov state model (MSM) and adaptive sampling techniques have also been used to study the transition states in protein oligomerization (Kelley et al. 2008; Jia et al. 2020).

Expanding horizons

Phase separation of proteins can lead to the formation of liquid droplets, colloidal suspensions, gelation, and solid aggregates. The main focus of the current review is on computational techniques associated with liquid-to-solid phase separation of proteins. Liquid-liquid phase separation (LLPS) driven by intermolecular interactions is an equally important phenomenon. LLPS is important for the formation of several biomolecular condensates, which are essential for cellular and nuclear functions. Understanding the mechanism and identification of proteins capable of liquid-liquid phase separation would help in understanding complex biological processes (Boeynaems et al. 2018). Choi et al. (2019) developed a lattice model–based simulation engine for exploring the phase separation of proteins. In this model, a protein molecule is modelled as 'stickers separated by spacer regions' to represent the regions that form inter-chain interactions. However, there are still a couple of open questions: (i) how does a cell control phase separation? and (ii) what decides the nature of the separated phase? These challenges offer new avenues for future research. For example, a unified computational model to predict both solid and liquid phase separation of proteins and peptides would help us understand cellular regulation and the mechanisms of biocondensate formation.

In an alternate direction, Mishra et al. (2018) applied the predictions from a protein aggregation prediction tool, AGGRESCAN, to screen native structures of proteins and their applications in protein tertiary structure prediction. Further, the design of peptide inhibitors, which selectively bind amyloid fibrils and fibril-forming proteins, is an active area of research (Lu et al. 2019; Seidler et al. 2019). These peptides bind to protofibrils and oligomers of amyloidogenic proteins to mitigate the protein aggregation in neurodegenerative diseases. In silico tools to predict the self-assembly of peptides have a wide range of applications. Several studies have shown the bactericidal activity of self-assembling peptides through the disruption of biofilm and cell membrane (Khodaparast et al. 2018; Lombardi et al. 2019; Tucker et al. 2018). The development of in silico tools for the designing and screening of such antimicrobial peptides could accelerate and widen the field. Peptide self-assembly has also been widely studied for its structural properties as a drug carrier and a scaffold in tissue engineering and constructing synthetic nanomaterials (Esteras-Chopo et al. 2005; Gallardo et al. 2016; Gupta et al. 2020; Hauser et al. 2014; Knowles and Mezzenga 2016). These diverging fields of protein aggregation provide scope and new horizon for the development of in silico tools.

Conclusions and future directions

Protein aggregation is a multidimensional phenomenon that involves diverse considerations such as stability of the native state, total aggregation propensity of the protein sequence, presence of aggregation-prone regions and gatekeeping residues, and environmental conditions such as concentration, pH, and ionic strength of protein solutions. Protein deposits such as tangles and plaques have been found in a diverse range of human pathologies with an undebatable association to the disease propagation itself. Predictions of the aggregation-prone region, propensity and aggregation rate of a protein sequence provide insights into its inherent tendency to drive intermolecular interactions and amyloid fibril formation. MD simulations have been instrumental in studying the protein oligomer formation and stability and peptide aggregation.

Peptide assembly and nanostructure formation are concentration-dependent and kinetically controlled phenomena. Studies have shown that external conditions could vary the nature and structure of aggregates. The methods currently available to predict the aggregation kinetics are still in the nascent stages. However, with the increase in experimental data, it may be possible to develop reliable next-generation kinetics models using large datasets. The inclusion of complex features such as pH, temperature, buffer, protein or ionic concentration, and agitating condition in computational models could help predict the rates of aggregation with greater accuracy. Similar limitations are also applicable to APR prediction tools. Most APR prediction tools assume the protein of interest exists predominately in the unfolded state and do not interact with other biomolecules. In contrast, cellular environments are highly crowded and only a small fraction of proteins are unfolded. For example, studies have shown cross-seeding, where a fibril fragment of a protein chain initiates amyloid formation of another (Ren et al. 2019). The phenomenon of cross-seeding is highly specific and mostly unidirectional.

In addition to the nucleation of aggregates, it is also important to understand propagation of the aggregation process. For example, what drives amyloid fibril polymorphism? Polymorphism in amyloid fibrils refers to the multiplicity of amyloid fibrillary structures formed by a given amyloidogenic peptide or protein. Researchers attribute polymorphism to the fibrillation kinetics and external conditions that influence the aggregation process. Addressing the polymorphism in zipper and protofibril structure predictions would pave ways for predicting and understanding complex nanostructure formations, which in turn could be useful for the design of novel biomaterials.

Understanding the molecular interactions that drive complex phenomena such as nucleation, polymorphism of amyloid fibrils, cross-seeding, etc. is essential to fully understand protein aggregations. Currently available experimental techniques are capable of playing only a limited role in this regard.

With the availability of increased computing power, multiscale molecular dynamics simulations are proving invaluable in elucidating these interactions. Exploiting recent advancements in sampling techniques, coarse-grained models, polarizable force fields, and constant-pH simulations shall enhance our understanding of the molecular events that determine the fate of protein aggregation.

Acknowledgements We thank the Bioinformatics Infrastructure Facility, Department of Biotechnology, and the Indian Institute of Technology Madras for computational facilities and the Ministry of Human Resource and Development (MHRD) for HTRA scholarship to PR. We thank WALTZ developers for sharing the executable.

References

Agrawal NJ, Helk B, Kumar S et al (2016) Computational tool for the early screening of monoclonal antibodies for their viscosities. MAbs 8:43–48. https://doi.org/10.1080/19420862.2015.1099773

Ahmed AB, Znassi N, Château M-T, Kajava AV (2015) A structure-based approach to predict predisposition to amyloidosis. Alzheimers Dement 11:681–690. https://doi.org/10.1016/j.jalz.2014.06.007

Angarica VE, Angulo A, Giner A et al (2014) PrionScan: an online database of predicted prion domains in complete proteomes. BMC Genomics 15:102. https://doi.org/10.1186/1471-2164-15-102

Astbury WT, Dickinson S, Bailey K (1935) The X-ray interpretation of denaturation and the structure of the seed globulins. Biochem J 29: 2351–2360.1. https://doi.org/10.1042/bj0292351

Azriel R, Gazit E (2001) Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. J Biol Chem 276:34156–34161. https://doi.org/10.1074/jbc.M102883200

Barducci A, Chelli R, Procacci P et al (2006) Metadynamics simulation of prion protein: β-structure stability and the early stages of misfolding. J Am Chem Soc 128:2705–2710. https://doi.org/10.1021/ja057076l

Beerten J, Van Durme J, Gallardo R, Capriotti E, Serpell L, Rousseau F, Schymkowitz J (2015) WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. Bioinformatics 31(10):1698–1700

Bellesia G, Shea J-E (2007) Self-assembly of β-sheet forming peptides into chiral fibrillar aggregates. J Chem Phys 126:245104. https://doi.org/10.1063/1.2739547

Bellesia G, Shea J-E (2009) Effect of β-sheet propensity on peptide aggregation. J Chem Phys 130:145103. https://doi.org/10.1063/1.3108461

Belli M, Ramazzotti M, Chiti F (2011) Prediction of amyloid aggregation in vivo. EMBO Rep 12:657–663. https://doi.org/10.1038/embor.2011.116

Benson MD, Buxbaum JN, Eisenberg DS et al (2018) Amyloid nomenclature 2018: recommendations by the International Society of Amyloidosis (ISA) nomenclature committee. Amyloid 25:215–219. https://doi.org/10.1080/13506129.2018.1549825

Berhanu WM, Masunov AE (2014) Full length amylin oligomer aggregation: insights from molecular dynamics simulations and implications for design of aggregation inhibitors. J Biomol Struct Dyn 32: 1651–1669. https://doi.org/10.1080/07391102.2013.832635

Bodi K, Prokaeva T, Spencer B et al (2009) AL-Base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. Amyloid 16:1–8. https://doi.org/10.1080/13506120802676781

Boeynaems S, Alberti S, Fawzi NL et al (2018) Protein phase separation: a new phase in cell biology. Trends Cell Biol 28:420–435. https://doi.org/10.1016/j.tcb.2018.02.004

Bondarev SA, Bondareva OV, Zhouravleva GA, Kajava AV (2018) BetaSerpentine: a bioinformatics tool for reconstruction of amyloid structures. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx629

Brito R, Damas A, Saraiva M (2003) Amyloid formation by transthyretin: from protein stability to protein aggregation. Curr Med Chem Endocr Metab Agents 3:349–360. https://doi.org/10.2174/1568013033483230

Brown AM, Bevan DR (2016) Molecular dynamics simulations of amyloid β-peptide (1-42): tetramer formation and membrane interactions. Biophys J 111:937–949. https://doi.org/10.1016/j.bpj.2016.08.001

Brudar S, Hribar-Lee B (2019) The role of buffers in wild-type HEWL amyloid fibril formation mechanism. Biomolecules 9:65. https://doi.org/10.3390/biom9020065

Bryan AW Jr, Menke M, Cowen LJ, Lindquist SL, Berger B (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol 5(3):e1000333. https://doi.org/10.1371/journal.pcbi.1000333

Bryan AW Jr, O'Donnell CW, Menke M, Cowen LJ, Lindquist S, Berger B (2012) STITCHER: dynamic assembly of likely amyloid and prion β-structures from secondary structure predictions. Proteins 80(2):410–420. https://doi.org/10.1002/prot.23203

Buck PM, Kumar S, Singh SK (2013) Insights into the potential aggregation liabilities of the b12 Fab fragment via elevated temperature molecular dynamics. Protein Eng Des Sel 26:195–206. https://doi.org/10.1093/protein/gzs099

Buck PM, Chaudhri A, Kumar S, Singh SK (2015) Highly viscous antibody solutions are a consequence of network formation caused by domain − domain electrostatic complementarities: insights from coarse-grained simulations. https://doi.org/10.1021/mp500485w

Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M (2017) Amyloidogenic motifs revealed by n-gram analysis. Sci Rep 7(1):12961. https://doi.org/10.1038/s41598-017-13210-9

Calamai M, Taddei N, Stefani M, Ramponi G, Chiti F (2003) Relative influence of hydrophobicity and net charge in the aggregation of two homologous proteins. Biochemistry 42(51):15078–15083

Carballo-Pacheco M, Strodel B (2016) Advances in the simulation of protein aggregation at the atomistic scale. J Phys Chem B 120: 2991–2999. https://doi.org/10.1021/acs.jpcb.6b00059

Cecchini M, Curcio R, Pappalardo M, Melki R, Caflisch A (2006) A molecular dynamics approach to the structural characterization of amyloid aggregation. J Mol Biol 357(4):1306–1321. https://doi.org/10.1016/j.jmb.2006.01.009

Chennamsetty N, Voynov V, Kayser V et al (2009) Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci U S A 106: 11937–11942. https://doi.org/10.1073/pnas.0904191106

Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75:333–366. https://doi.org/10.1146/annurev.biochem.75.101304.123901

Chiti F, Calamai M, Taddei N et al (2002a) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. Proc Natl Acad Sci U S A 99:16419–16426. https://doi.org/10.1073/pnas.212527999

Chiti F, Taddei N, Baroni F, Capanni C, Stefani M, Ramponi G, Dobson CM (2002b) Kinetic partitioning of protein folding and aggregation. Nat Struct Biol 9(2):137–143

Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424(6950):805–808

Choi J-M, Dar F, Pappu RV (2019) LASSI: a lattice model for simulating phase transitions of multivalent proteins. PLoS Comput Biol 15: e1007028. https://doi.org/10.1371/journal.pcbi.1007028

Conchillo-Solé O, de Groot NS, Avilés FX et al (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinform 8:65. https://doi.org/10.1186/1471-2105-8-65

Dogan A (2017) Amyloidosis: insights from proteomics. Annu Rev Pathol Mech Dis 12:277–304. https://doi.org/10.1146/annurev-pathol-052016-100200

DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol 341(5):1317–1326

Emily M, Talvas A, Delamarche C (2013) MetAmyl: a METa-predictor for AMYLoid proteins. PLoS One 8(11):e79722. https://doi.org/10.1371/journal.pone.0079722

Esteras-Chopo A, Serrano L, López de la Paz M (2005) The amyloid stretch hypothesis: recruiting proteins toward the dark side. Proc Natl Acad Sci U S A 102:16672–16677. https://doi.org/10.1073/pnas.0505905102

Família C, Dennison SR, Quintas A, Phoenix DA (2015) Prediction of peptide and protein propensity for amyloid formation. PLoS One 10(8):e0134679. https://doi.org/10.1371/journal.pone.0134679

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

Fink AL (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid. Fold Des 3(1):R9–R23

Frederix PWJM, Ulijn RV, Hunt NT, Tuttle T (2011) Virtual screening for dipeptide aggregation: toward predictive tools for peptide self-assembly. J Phys Chem Lett 2:2380–2384. https://doi.org/10.1021/jz2010573

Frederix PWJM, Scott GG, Abul-Haija YM et al (2015) Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels. Nat Chem 7:30–37. https://doi.org/10.1038/nchem.2122

Gallardo R, Ramakers M, De Smet F et al (2016) De novo design of a biologically active amyloid. Science. 354:6313. https://doi.org/10.1126/science.aah4949

Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26:326–332. https://doi.org/10.1093/bioinformatics/btp691

Gasior P, Kotulska M (2014) FISH Amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. BMC Bioinform 15:54. https://doi.org/10.1186/1471-2105-15-54

Gazit E (2002) A possible role for π-stacking in the self-assembly of amyloid fibrils. FASEB J 16(1). https://doi.org/10.1096/fj.01-0442hyp

Gertz M. A. (2018) Annual clinical updates in hematological malignancies : a continuing medical education series immunoglobulin light chain amyloidosis : 2018 update on diagnosis , prognosis , and treatment. 1169–1180. https://doi.org/10.1002/ajh.25149

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE (2018) UCSF ChimeraX: meeting modern challenges in visualization and analysis. Protein Sci 27(1):14–25

Goldschmidt L, Teng PK, Riek R, Eisenberg D (2010) Identifying the amylome, proteins capable of forming amyloid-like fibrils. Proc Natl Acad Sci 107(8):3487–3492

Green AA, Hughes WL (1955) Protein fractionation on the basis of solubility in aqueous solutions of salts and organic solvents. Methods Enzymol 1:67–90. https://doi.org/10.1016/0076-6879(55)01014-8

Gsponer J, Haberthur U, Caflisch A (2003) The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. Proceedings of the National Academy of Sciences 100 (9): 5154–5159. https://doi.org/10.1073/pnas.0835307100

Guo C, Côté S, Mousseau N, Wei G (2015) Distinct helix propensities and membrane interactions of human and rat IAPP 1–19 monomers in anionic lipid bilayers. J Phys Chem B 119:3366–3376. https://doi.org/10.1021/jp5111357

Gupta S, Singh I, Sharma AK, Kumar P (2020) Ultrashort peptide self-assembly: front-runners to transport drug and gene cargos. Front Bioeng Biotechnol 8. https://doi.org/10.3389/fbioe.2020.00504

Hamodrakas SJ, Liappa C, Iconomidou VA (2007) Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. Int J Biol Macromol 41(3):295–300. https://doi.org/10.1016/j.ijbiomac.2007.03.008

Hauser CA, Maurer-Stroh S, Martins IC (2014) Amyloid-based nanosensors and nanodevices. Chem Soc Rev 43(15):5326–5345. https://doi.org/10.1039/c4cs00082j

Hilbich C, Kisters-Woike B, Reed J, Masters CL, Beyreuther K (1992) Substitutions of hydrophobic amino acids reduce the amyloidogenicity of Alzheimer's disease βA4 peptides. J Mol Biol 228(2):460–473

Hirota N, Edskes H, Hall D (2019) Unified theoretical description of the kinetics of protein aggregation. Biophys Rev 11(2):191–208

Hortschansky P, Schroeckh V, Christopeit T, Zandomeneghi G, Fändrich M (2005) The aggregation kinetics of Alzheimer's β-amyloid peptide is controlled by stochastic nucleation. Protein Sci 14(7):1753–1759

Iadanza MG, Jackson MP, Hewitt EW et al (2018) A new era for understanding amyloid structures and disease. Nat Rev Mol Cell Biol 19: 755–773. https://doi.org/10.1038/s41580-018-0060-8

Itoh SG, Okumura H (2013) Hamiltonian replica-permutation method and its applications to an alanine dipeptide and amyloid-β (29–42) peptides. J Comput Chem 34(29):2493–2497

Itoh SG, Okumura H (2016) Oligomer formation of amyloid-β (29–42) from its monomers using the Hamiltonian replica-permutation molecular dynamics simulation. J Phys Chem B 120(27):6555–6561

Jain T, Sun T, Durand S et al (2017) Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci 114:944–949. https://doi.org/10.1073/pnas.1616408114

Jia Z, Schmit JD, Chen J (2020) Amyloid assembly is dominated by misregistered kinetic traps on an unbiased energy landscape. Proc Natl Acad Sci 117:10322–10328. https://doi.org/10.1073/pnas.1911153117

Karandur D, Wong KY, Pettitt BM (2014) Solubility and aggregation of Gly5in water. J Phys Chem B 118(32):9565–9572. https://doi.org/10.1021/jp503358n

Ke PC, Zhou R, Serpell LC, Riek R, Knowles TPJ, Lashuel HA, Gazit E, Hamley IW, Davis TP, Fändrich M, Otzen DE, Chapman MR, Dobson CM, Eisenberg DS, Mezzenga R (2020) Half a century of amyloids: past, present and future. Chem Soc Rev 49(15):5473–5509. https://doi.org/10.1039/C9CS00199A

Kelley NW, Vishal V, Krafft GA, Pande VS (2008) Simulating oligomerization at experimental concentrations and long timescales: a Markov state model approach. J Chem Phys 129:214707. https://doi.org/10.1063/1.3010881

Khatua P, Bandyopadhyay S (2017) In silico studies of the early stages of aggregation of A β42 peptides. J Chem Sci 129:899–909. https://doi.org/10.1007/s12039-017-1306-2

Khodaparast L, Khodaparast L, Gallardo R et al (2018) Aggregating sequences that occur in many proteins constitute weak spots of bacterial proteostasis. Nat Commun 9:866. https://doi.org/10.1038/s41467-018-03131-0

Kim C, Choi J, Lee SJ, Welsh WJ, Yoon S (2009) NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. Nucleic Acids Res 37(Web Server issue):W469–W473. https://doi.org/10.1093/nar/gkp351

Kintzing JR, Filsinger Interrante MV, Cochran JR (2016) Emerging strategies for developing next-generation protein therapeutics for cancer treatment. Trends Pharmacol Sci 37:993–1008. https://doi.org/10.1016/j.tips.2016.10.005

Knowles TPJ, Buehler MJ (2011) Nanomechanics of functional and pathological amyloid materials. Nat Nanotechnol 6:469–479. https://doi.org/10.1038/nnano.2011.102

Knowles TPJ, Mezzenga R (2016) Amyloid fibrils as building blocks for natural and artificial functional materials. Adv Mater 28:6546–6561. https://doi.org/10.1002/adma.201505961

Kumar S, Roffi K, Tomar, Dheeraj S et al (2018) Rational optimization of a monoclonal antibody for simultaneous improvements in its solution properties and biological activity. *Protein Eng Des Sel* **31**:313–325

Kumar V, Wahiduzzaman PA et al (2019) Exploring the aggregation-prone regions from structural domains of human TDP-43. Biochim Biophys Acta, Proteins Proteomics 1867:286–296. https://doi.org/10.1016/j.bbapap.2018.10.008

Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S (2019) Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. Nucleic Acids Res 47(W1):W300–W307. https://doi.org/10.1093/nar/gkz321

Kyle RA, Bayrd ED (1975) Amyloidosis: review of 236 cases. Medicine (Baltimore) 54:271–299. https://doi.org/10.1097/00005792-197507000-00001

Lagassé HAD, Alexaki A, Simhadri VL et al (2017) Recent advances in (therapeutic protein) drug development. F1000Research 6:113. https://doi.org/10.12688/f1000research.9970.1

Larini L, Shea J-E (2012) Role of β-hairpin formation in aggregation: the self-assembly of the amyloid-β(25−35) peptide. Biophys J 103: 576–586

Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL (2012) Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. J Pharm Sci 101(1):102–115. https://doi.org/10.1002/jps.22758

Leader B, Baca QJ, Golan DE (2008) Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov 7:21–39. https://doi.org/10.1038/nrd2399

Leem J, Dunbar J, Georges G, Shi J, Deane CM, (2016) ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. mAbs 8(7):1259–1268

Li W, Prabakaran P, Chen W et al (2016) Antibody aggregation: insights from sequence and structure. Antibodies 5:19. https://doi.org/10.3390/antib5030019

Liaw C, Tung CW, Ho SY (2013) Prediction and analysis of antibody amyloidogenesis from sequences. PLoS One 8(1):e53235. https://doi.org/10.1371/journal.pone.0053235

Lombardi L, Shi Y, Falanga A et al (2019) Enhancing the potency of antimicrobial peptides through molecular engineering and self-assembly. Biomacromolecules 20:1362–1374. https://doi.org/10.1021/acs.biomac.8b01740

López de la Paz M, Serrano L (2004) Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci U S A 101(1):87–92. https://doi.org/10.1073/pnas.2634884100

López De La Paz M, De Mori GMS, Serrano L, Colombo G (2005) Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. J Mol Biol 349:583–596. https://doi.org/10.1016/j.jmb.2005.03.081

Louros N, Konstantoulea K, De Vleeschouwer M, Ramakers M, Schymkowitz J, Rousseau F (2020) WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. Nucleic Acids Res 48(D1):D389–D393. https://doi.org/10.1093/nar/gkz758

Lu J, Cao Q, Wang C et al (2019) Structure-based peptide inhibitor design of amyloid-β aggregation. Front Mol Neurosci 12:54. https://doi.org/10.3389/fnmol.2019.00054

Ma B, Nussinov R (2002a) Molecular dynamics simulations of alanine rich β-sheet oligomers: insight into amyloid formation. Protein Sci 11:2335–2350. https://doi.org/10.1110/ps.4270102

Ma B, Nussinov R (2002b) 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: 14126–14131. https://doi.org/10.1073/pnas.212206899

Magno A, Caflisch A, Pellarin R (2010) Crowding effects on amyloid aggregation kinetics. J Phys Chem Lett 1:3027–3032. https://doi.org/10.1021/jz100967z

Marchut AJ, Hall CK (2006) Side-chain interactions determine amyloid formation by model polyglutamine peptides in molecular dynamics simulations. Biophys J 90(12):4574–4584. https://doi.org/10.1529/biophysj.105.079269

Marchut AJ, Hall CK (2007) Effects of chain length on the aggregation of model polyglutamine peptides: molecular dynamics simulations. Proteins Struct Funct Genet. https://doi.org/10.1002/prot.21132

Matthes D, Gapsys V, Daebel V, de Groot BL (2011) Mapping the conformational dynamics and pathways of spontaneous steric zipper peptide oligomerization. PLoS One 6:e19129. https://doi.org/10.1371/journal.pone.0019129

Matthes D, Gapsys V, De Groot BL (2012) Driving forces and structural determinants of steric zipper peptide oligomer formation elucidated by atomistic simulations. J Mol Biol 421:390–416. https://doi.org/10.1016/j.jmb.2012.02.004

Maurer-Stroh S, Debulpaep M, Kuemmerer N et al (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7:237–242. https://doi.org/10.1038/nmeth.1432

Meric G, Robinson AS, Roberts CJ (2017) Driving forces for nonnative protein aggregation and approaches to predict aggregation-prone regions. Annu Rev Chem Biomol Eng 8:139–159. https://doi.org/10.1146/annurev-chembioeng-060816-101404

Mishra A, Ranganathan S, Jayaram B, Sattar A (2018) Role of solvent accessibility for aggregation-prone patches in protein folding. Sci Rep 8:12896. https://doi.org/10.1038/s41598-018-31289-6

Morel B, Varela L, Azuaga AI, Conejero-Lara F (2010) Environmental conditions affect the kinetics of nucleation of amyloid fibrils and determine their morphology. Biophys J 99(11):3801–3810

Morris AM, Watzky MA, Finke RG (2009) Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature. Biochim Biophys Acta (BBA)-Proteins Proteom 1794(3):375–397

Morriss-Andrews A, Shea JE (2014) Simulations of protein aggregation: insights from atomistic and coarse-grained models. J Phys Chem Lett 5:1899–1908. https://doi.org/10.1021/jz5006847

Morriss-Andrews A, Shea J-E (2015) Computational studies of protein aggregation: methods and applications. Annu Rev Phys Chem 66: 643–666. https://doi.org/10.1146/annurev-physchem-040513-103738

Muñoz V, Serrano L (1994) Elucidating the folding problem of helical peptides using empirical parameters. Nat Struct Biol 1(6):399–409

Nelson R, Sawaya MR, Balbirnie M et al (2005) Structure of the cross-beta spine of amyloid-like fibrils. Nature 435:773–778. https://doi.org/10.1038/nature03680

Nguyen HD, Hall CK (2004a) Phase diagrams describing fibrillization by polyalanine peptides. Biophys J 87:4122–4134. https://doi.org/10.1529/biophysj.104.047159

Nguyen HD, Hall CK (2004b) Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. Proc Natl Acad Sci U S A 101:16180–16185. https://doi.org/10.1073/pnas.0407273101

Nguyen HD, Hall CK (2005) Kinetics of fibril formation by polyalanine peptides. J Biol Chem 280:9074–9082. https://doi.org/10.1074/jbc.M407338200

Nguyen HD, Hall CK (2006) Spontaneous fibril formation by polyalanines; discontinuous molecular dynamics simulations. J Am Chem Soc 128:1890–1901. https://doi.org/10.1021/ja0539140

Nichols P, Li L, Kumar S et al (2015) Rational design of viscosity reducing mutants of a monoclonal antibody: hydrophobic versus electrostatic inter-molecular interactions. MAbs 7:212–230. https://doi.org/10.4161/19420862.2014.985504

O'Donnell CW, Waldispühl J, Lis M, Halfmann R, Devadas S, Lindquist S, Berger B (2011) A method for probing the mutational landscape of amyloid structure. Bioinformatics 27(13):i34–i42. https://doi.org/10.1093/bioinformatics/btr238

Orlando G, Silva A, Macedo-Ribeiro S, Raimondi D, Vranken W (2020) Accurate prediction of protein beta-aggregation with generalized statistical potentials. Bioinformatics 36(7):2076–2081

Ow SY, Dunstan DE (2013) The effect of concentration, temperature and stirring on hen egg white lysozyme amyloid formation. Soft Matter 9(40):9692–9701

Paparcone R, Cranford SW, Buehler MJ (2011) Self-folding and aggregation of amyloid nanofibrils. Nanoscale 3:1748–1755. https://doi.org/10.1039/c0nr00840k

Pawlicki S, Le Béchec A, Delamarche C (2008) AMYPdb: a database dedicated to amyloid precursor proteins. BMC Bioinform 9(1):273

Peng S, Ding F, Urbanc B et al (2004) Discrete molecular dynamics simulations of peptide aggregation. Phys Rev E Stat Nonlinear Soft Matter Phys 69:041908. https://doi.org/10.1103/PhysRevE.69.041908

Polanco C, Samaniego JL, Uversky VN, Castañón-González JA, Buhse T, Leopold-Sordo M, ... Arias-Estrada M (2015) Identification of proteins associated with amyloidosis by polarity index method. Acta Biochim Polonica 62(1)

Prabakaran R, Rawat P, Kumar S, Gromiha MM (2020) ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins. J Mol Biol (in press). https://doi.org/10.1016/j.jmb.2020.11.006

Priya SB, Gromiha MM (2019) Structural insights into the aggregation mechanism of huntingtin exon 1 protein fragment with different polyQ-lengths. J Cell Biochem 120(6):10519–10529. https://doi.org/10.1002/jcb.28338

Ramírez-Alvarado M, Merkel JS, Regan L (2000) A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. Proc Natl Acad Sci 97(16):8979–8984

Rawat P, Kumar S, Gromiha MM (2018) An in-silico method for identifying aggregation rate enhancer and mitigator mutations in proteins. Int J Biol Macromol 118:1157–1167

Rawat P, Prabakaran R, Sakthivel R, Mary Thangakani A, Kumar S, Gromiha MM (2020a) CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. Amyloid 27(2): 128–133

Rawat P, Prabakaran R, Kumar S, Gromiha MM (2020b) AggreRATE-Pred: a mathematical model for the prediction of change in aggregation rate upon point mutation. Bioinformatics 36(5):1439–1444

Ren B, Zhang Y, Zhang M, Liu Y, Zhang D, Gong X, Feng Z, Tang J, Chang Y, Zheng J (2019) Fundamentals of cross-seeding of amyloid proteins: an introduction. J Mater Chem B 7(46):7267–7282

Roberts CJ (2014) Protein aggregation and its impact on product quality. Curr Opin Biotechnol 30:211–217. https://doi.org/10.1016/j.copbio.2014.08.001

Sankar K, Krystek SR, Carl SM et al (2018) AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins Struct Funct Bioinforma 86:1147–1156. https://doi.org/10.1002/prot.25594

Sawaya MR, Sambashivan S, Nelson R et al (2007) Atomic structures of amyloid cross-β spines reveal varied steric zippers. Nature 447: 453–457. https://doi.org/10.1038/nature05695

Seidler PM, Boyer DR, Murray KA et al (2019) Structure-based inhibitors halt prion-like seeding by Alzheimer's disease-and tauopathy-derived brain tissue samples. J Biol Chem 294:16451–16464. https://doi.org/10.1074/jbc.RA119.009688

Shobana R, Pandaranayaka EP (2014) ProADD: a database on protein aggregation diseases. Bioinformation 10(6):390

Siepen JA, Westhead DR (2002) The fibril_one on-line database: mutations, experimental conditions, and trends associated with amyloid fibril formation. Protein Sci 11(7):1862–1866

Singh G, Brovchenko IV, Oleinikova A, Winter R (2008) Peptide aggregation in finite systems. Biophys J 95:3208–3221. https://doi.org/10.1529/biophysj.108.136226

Sipe JD, Cohen AS (2000) Review: History of the amyloid fibril. J Struct Biol 130:88–98. https://doi.org/10.1006/jsbi.2000.4221

Sunde M, Blake C (1997) The structure of amyloid fibrils by electron microscopy and x-ray diffraction. Adv Protein Chem 50:123–159. https://doi.org/10.1016/s0065-3233(08)60320-4

Swuec P, Lavatelli F, Tasaki M et al (2019) Cryo-EM structure of cardiac amyloid fibrils from an immunoglobulin light chain AL amyloidosis patient. Nat Commun 10:1269

Takács K, Varga B, Grolmusz V (2019) PDB _Amyloid: an extended live amyloid structure list from the PDB. FEBS Open Bio 9(1):185–190

Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev 37: 1395–1401. https://doi.org/10.1039/b706784b

Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2004) The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. Protein Sci 13(7):1939–1941

Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Sci 14(10):2723–2734. https://doi.org/10.1110/ps.051471205

Thangakani AM, Kumar S, Nagarajan R et al (2014) GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics 30:1983–1990. https://doi.org/10.1093/bioinformatics/btu167

Thangakani AM, Nagarajan R, Kumar S, Sakthivel R, Velmurugan D, Gromiha MM (2016) CPAD, curated protein aggregation database: a repository of manually curated experimental data on protein and peptide aggregation. PLoS One 11(4):e0152949

Thompson MJ, Sievers SA, Karanicolas J et al (2006) The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci 103:4074–4078. https://doi.org/10.1073/pnas.0511295103

Tian J, Wu N, Guo J, Fan Y (2009) Prediction of amyloid fibril-forming segments based on a support vector machine. BMC Bioinform 10 Suppl 1(Suppl 1):S45. https://doi.org/10.1186/1471-2105-10-S1-S45

Tiller KE, Li L, Kumar S et al (2017) Arginine mutations in antibody complementarity-determining regions display context-dependent affinity/specificity trade-offs. J Biol Chem 292:16638–16652. https://doi.org/10.1074/jbc.M117.783837

Tjernberg L, Hosia W, Bark N, Thyberg J, Johansson J (2002) Charge attraction and β propensity are necessary for amyloid fibril formation from tetrapeptides. J Biol Chem 277(45):43243–43246

Tran L, Ha-Duong T (2015) Exploring the Alzheimer amyloid-β peptide conformational ensemble: a review of molecular dynamics approaches. Peptides 69:86–91. https://doi.org/10.1016/j.peptides.2015.04.009

Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ (2013) A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. PLoS One 8(1):e54175. https://doi.org/10.1371/journal.pone.0054175

Tucker AT, Leonard SP, DuBois CD, Knauf GA, Cunningham AL, Wilke CO, Trent MS, Davies BW (2018) Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of Surface-Displayed Peptide Libraries. Cell 172(3):618.e13–628.e13. https://doi.org/10.1016/j.cell.2017.12.009

Usmani SS, Bedi G, Samuel JS et al (2017) THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS One 12:1–12. https://doi.org/10.1371/journal.pone.0181748

Varadi M, De Baets G, Vranken WF et al (2018) AmyPro: a database of proteins with validated amyloidogenic regions. Nucleic Acids Res 46:D387–D392. https://doi.org/10.1093/nar/gkx950

Ventura S, Zurdo J, Narayanan S et al (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. Proc Natl Acad Sci U S A 101:7258–7263. https://doi.org/10.1073/pnas.0308249101

Walsh I, Seno F, Tosatto SCE, Trovato A (2014) PASTA 2.0: An improved server for protein aggregation prediction. Nucleic Acids Res 42:301–307. https://doi.org/10.1093/nar/gku399

Wang X, Das TK, Singh SK, Kumar S (2009) Potential aggregation prone regions in biotherapeutics: a survey of commercial monoclonal antibodies. MAbs 1:254–267. https://doi.org/10.4161/mabs.1.3.8035

Wang Y, Bunce SJ, Radford SE, Wilson AJ, Auer S, Hall CK (2019) Thermodynamic phase diagram of amyloid-β (16–22) peptide. Proc Natl Acad Sci 116(6):2091–2096. https://doi.org/10.1073/pnas.1819592116

Wojciechowski JW, Kotulska M (2020) PATH - Prediction of Amyloidogenicity by Threading and Machine Learning. Sci Rep 10(1):7721. https://doi.org/10.1038/s41598-020-64270-3

Wozniak PP, Kotulska M (2015) AmyLoad: website dedicated to amyloidogenic protein fragments. Bioinformatics 31(20):3395–3397

Wu C, Lei H, Duan Y (2005) Elongation of ordered peptide aggregate of an amyloidogenic hexapeptide NFGAIL observed in molecular dynamics simulations with explicit solvent. J Am Chem Soc 127: 13530–13537. https://doi.org/10.1021/ja050767x

Yagi-Utsumi M, Yanaka S, Song C et al (2020) Characterization of amyloid β fibril formation under microgravity conditions. NPJ Micrograv 6:17. https://doi.org/10.1038/s41526-020-0107-y

Yang W, Tan P, Fu X, Hong L (2019) Prediction of amyloid aggregation rates by machine learning and feature selection. J Chem Phys 151(8):084106

Zambrano R, Jamroz M, Szczasiuk A et al (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. Nucleic Acids Res 43:W306–W313. https://doi.org/10.1093/nar/gkv359

Zanuy D, Nussinov R (2003) The sequence dependence of fiber organization. A comparative molecular dynamics study of the islet amyloid polypeptide segments 22-27 and 22-29. J Mol Biol 329:565–584. https://doi.org/10.1016/S0022-2836(03)00491-1

Zanuy D, Ma B, Nussinov R (2003) Short peptide amyloid organization: stabilities and conformations of the islet amyloid peptide NFGAIL. Biophys J 84:1884–1894. https://doi.org/10.1016/S0006-3495(03)74996-0

Zemla A, Venclovas Č, Fidelis K, Rost B (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins Struct Funct Genet 34(2):220–223. https://doi.org/10.1002/(SICI)1097-0134(19990201)34:2<220::AID-PROT7>3.0.CO;2-K

Zhang Z, Chen H, Lai L (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics (Oxford, England) 23(17):2218–2225. https://doi.org/10.1093/bioinformatics/btm325

Zheng W, Tsai MY, Chen M, Wolynes PG (2016) Exploring the aggregation free energy landscape of the amyloid-β protein (1-40). Proc Natl Acad Sci U S A 113(42):11835–11840. https://doi.org/10.1073/pnas.1612362113

Zibaee S, Makin OS, Goedert M, Serpell LC (2007) A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. Protein Sci 16(5):906–918. https://doi.org/10.1110/ps.062624507

Zurdo J (2013) Developability assessment as an early de-risking tool for biopharmaceutical development. Pharm Bioprocess 1:29–50. https://doi.org/10.4155/pbp.13.3