Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Exploring the relationships between protein sequence, structure and solubility

Kyle Trainor, Aron Broom and Elizabeth M Meiering

Aggregation can be thought of as a form of protein folding in which intermolecular associations lead to the formation of large, insoluble assemblies. Various types of aggregates can be differentiated by their internal structures and gross morphologies (e.g., fibrillar or amorphous), and the ability to accurately predict the likelihood of their formation by a given polypeptide is of great practical utility in the fields of biology (including the study of disease), biotechnology, and biomaterials research. Here we review aggregation/solubility prediction methods and selected applications thereof. The development of increasingly sophisticated methods that incorporate knowledge of conformations possibly adopted by aggregating polypeptide monomers and predict the internal structure of aggregates is improving the accuracy of the predictions and continually expanding the range of applications.

Address
Department of Chemistry, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada

Corresponding author: Meiering, Elizabeth M (meiering@uwaterloo.ca)

Current Opinion in Structural Biology 2017, **42**:136–146

This review comes from a themed issue on **Folding and binding**

Edited by **Jane Clarke** and **Rohit V Pappu**

For a complete overview see the Issue and the Editorial

Available online 2nd February 2017

http://dx.doi.org/10.1016/j.sbi.2017.01.004

Introduction

Aggregation can be defined as protein self-association that results in large, insoluble assemblies. The accurate prediction of protein aggregation and solubility is of significant practical utility in the production of recombinant proteins for research and biotechnological purposes [1], exploration of the properties of natural and engineered protein assemblies [2], the formulation of biopharmaceuticals [3], and in the context of protein misfolding-related disease [4]. Here we review recent developments in the field of aggregation/solubility prediction, including computational methods, their bases and selected applications. The successes and limitations of these methods reveal much about the extent of our understanding of the relationships between protein sequence, structure, and solubility. Many methods have proven capable of identifying aggregation prone regions (APRs). A few also consider the exposure of APRs in near-native structures or the predicted thermodynamic stability of the native state; however, insights into the mechanisms and consequences of APR exposure due to full or partial unfolding that are general enough to be incorporated into predictive methods remain elusive.

Although protein aggregation is a complicated phenomenon, many fundamental aspects are understood. It occurs as a result of the same influences responsible for 'normal' protein folding (i.e., to the soluble native state) [5]; thus, aggregation can be thought of as folding to an alternate state with features common to almost all thermodynamically stable protein conformations: relative compactness, substantial desolvation of hydrophobic side chains, and the satisfaction of many potential hydrogen bonds. A classic example is the amyloid fibril, a type of protein aggregate with a high degree of long-range order, extensive intermolecular (β-sheet) hydrogen bonding, and a well-packed core [6]. More generally, the degree of long-range order within aggregates may vary, and the conformations adopted by the constituent monomers may include native, native-like, and/or non-native secondary/tertiary structure. Gross aggregate morphologies range from fibrillar to amorphous, and a given polypeptide can form structurally distinct aggregates that may propagate in a prion-like fashion [7]. Furthermore, *in vivo* aggregation can involve additional complications such as macromolecular crowding and the activity of chaperones or proteases [8]; the impacts of such factors may be implicitly incorporated into predictive methods based on data from *in vivo* experiments.

The development of protein aggregation/solubility prediction methods has been underway for several decades, and continues to be an active area of research (Table 1). Over this time, many methodological improvements have been driven by the accumulation of fundamental insights into the thermodynamics [9] and kinetics [10•] of aggregation, the realization that short sequence segments may determine aggregation propensity [11,12], and advances in the modeling and simulation of dynamic aggregation-prone surface exposure [13•,14,15••]. Progress in the development of increasingly sophisticated prediction methods is described below (Table 1), along with selected applications that illustrate the utility of the methods (Table 2).

Collectively, the aggregation/solubility prediction methods reviewed here have been applied to diverse problems,

Table 1

Solubility/aggregation prediction algorithms[a]

Method	Year	Basis of prediction	Ref.
Amino acid composition-based algorithms			
Chiti-Dobson	2003	The natural logarithm of the ratio between mutant and wild-type aggregation rates predicted by the weighted sum of change in hydrophobicity, change in secondary structure propensity, and change in net charge	[10•]
Price-Hunt	2011	Statistical analysis of high-throughput expression of proteins by the Northeast Structural Genomics Consortium	[31•]
SCM	2012	Dipeptide solubility scoring matrix	[28]
PROSO II	2012	Two-layer architecture in which the output of a Parzen window model for sequence similarity and an amino acid composition logistic regression classifier feed into a second logistic regression classifier	[18]
Samak-Wang	2012	SVM and RF classifiers trained on the data from [17] and used in combination	[19]
CCSOL	2012	SVM trained to discriminate between soluble and insoluble proteins in the data from [17]	[20]
Niu-Li	2014	SVM solubility predictions based on pseudo amino acid composition models with features including CGRs and Shannon entropy	[36]
PON-Sol	2016	Effect of amino acid substitutions predicted using a three-class (solubility increasing, decreasing, or unchanged), two-layer RF classifier	[29]
Sliding window/pattern-based algorithms			
TANGO	2004	Percent occupancy of major conformational states (including β-aggregate) predicted for each residue	[23]
3D Profile (ZipperDB)	2006	Compatibility of sequence segments with the conformation adopted by the NNQQNY hexapeptide in cross-β protofibrils	[22]
PASTA	2007	Statistical analysis of residue pairings between adjacent β-strands in known structures	[48]
AGGRESCAN	2007	Sliding window average of aggregation propensity scores for amino acids derived from measurements of intracellular aggregation by Aβ42 mutants	[21]
Zyggregator	2008	Sliding window average of aggregation propensity scores adjusted for gatekeepers and alternating patterns of hydrophobic and hydrophilic residues	[25]
FoldAmyloid	2010	Sliding window average of amino acid packing density and hydrogen bond probability scores	[24]
Waltz	2010	PSSM derived from amyloidogenic hexapeptides, physicochemical properties, and structural modeling using amyloid backbone structures	[51]
AmyloidMutants	2011	Supersecondary structure prediction enables discrimination between different amyloid configurations, with optional mutational analysis	[56]
ESPRESSO	2013	Binary classification of sequences using predicted (secondary) structural properties and sequence pattern-based methods	[40]
PASTA 2.0	2014	Statistical analysis of residue pairings between adjacent β-strands in known structures	[38]
FISH Amyloid	2014	Binary classification of sequence segments using a discriminative pattern of site-specific co-occurrences of residue pairs in known amyloidogenic hexapeptides	[39]
CamSol	2015	Sliding window average of aggregation propensity scores adjusted for gatekeepers and alternating patterns of hydrophobic and hydrophilic residues	[26]
Tertiary/quaternary structure-based algorithms			
SAP	2009	Effective dynamically exposed hydrophobic surface patches determined by structural analysis and short MD simulations	[13•]
Chan-Warwicker	2013	Correlation between positively charged surface patches and insoluble expression, particularly when the patch is enriched in arginine relative to lysine	[72•]
CamSol	2015	Zyggregator-like sequence-based predictions [25] projected onto a 3D structure and adjusted for solvent exposure and the influence of other residues within an 8 Åradius	[26]
AGGRESCAN3D	2015	AGGRESCAN sequence-based predictions [21] projected onto a 3D structure and adjusted for solvent exposure and the influence of other residues within a 10 Åradius; optional simulation of dynamic exposure using CABS-flex [70•]	[14]
Schaller-Middleberg	2015	Parameters recorded during MD simulations of thermal unfolding at 498 K used as input for an SVM classifier	[15••]

[a] Developed or applied (in published literature) between 2011 and 2016 inclusive.

such as computationally screening biotherapeutics for solubility [16], exploring the determinants of solubility in the *E. coli* proteome [17–20], and enhancing the solubility of aggregation-prone proteins through mutations [13•,14,21–26] and strategic glycosylation [27]. In the sections that follow, we group these methods into three broad categories: first, statistical analyses and machine learning algorithms that abstract aggregation-related features from the amino acid sequences of proteins with known aggregation propensities, second, 'sliding window'