Protein aggregation has been increasingly recognized as a problem limiting the efficacy and shelf life of protein therapeutics and as an indicator and cause of numerous disease states. Elucidating the molecular mechanisms behind aggregation has become a central focus of investigation in order to improve therapeutics and to understand the relationship between aggregate formation and cellular toxicity in protein misfolding diseases. Innovations in analysis techniques, particularly of solid-state materials, and computational molecular modeling approaches have provided higher resolution information about the structure of aggregates as well as key insights into the mechanisms of aggregate formation. These breakthroughs, coupled with understanding gained from solution experiments and biological systems, have just begun to enable strategies to combat aggregation, including the design and evaluation of peptides and small molecules that inhibit the growth or that facilitate the dissociation of aggregates. This chapter describes the fundamental properties of proteins and the current understanding of underlying mechanisms that influence native folding and the formation of aggregates.

1.1 THE PROBLEM OF PROTEIN AGGREGATION

Protein aggregation has significant influence in the pathology, onset, and progression of most, if not all, misfolding diseases. Over 40 human diseases have been linked to aggregation of a specific protein, including hemoglobin in sickle cell anemia, the widely recognized Aβ peptides in Alzheimer’s disease, the PrP prion protein in Creutzfeldt–Jakob’s and related diseases, expanded polyglutamine tracts in Huntington’s disease, amylin-induced β-cell death in
diabetes, and \(\alpha\)-synuclein in Parkinson’s disease.\(^1\) Moreover, studies of non-disease-associated proteins *in vitro* show that aggregates and amyloid fibers can be induced to occur from almost any protein, suggesting it is a ubiquitous phenomenon reflecting a common mechanism.\(^2\) Therapeutic proteins used to treat various diseases can also produce ill effects when aggregates are present, in some cases contributing to amyloid plaque formation *in vivo*.\(^3,4\) Aggregates have been observed to form in therapeutic proteins during purification and storage, and the administration of proteins containing aggregates has been shown to stimulate immune responses, causing effects ranging from mild skin irritation to anaphylaxis.\(^5,6\) As such, major efforts are underway to stabilize therapeutic proteins against aggregation. Thus, the goal of understanding the fundamental properties of proteins that contribute to aggregation and the mechanisms by which they aggregate is of critical importance for determining how to prevent and treat numerous diseases.

*In vivo* protein aggregation appears to be an ever-present problem caused by thermal fluctuations and chemical changes that disrupt the physical structure of these delicate molecules. Consequently, cells have evolved several mechanisms by which they prevent aggregates from interfering with normal function.\(^7\) Improperly folded proteins are removed from cells before they can initiate aggregation by being degraded into smaller peptides via the proteosome or lysosomal enzymes. Alternatively, intracellular proteins can be refolded to their native conformation by interaction with chaperone proteins, which are often expressed at elevated levels in response to thermal (heat shock) or chemical stress. Chaperones bind to hydrophobic patches on misfolded proteins and use an energy-dependent process to alter their conformation, therein providing the protein with additional attempts to find its native fold.\(^8\) When the capacity of the aforementioned machinery is exceeded, aggregates may form,\(^9\) as is often observed in recombinant expression systems. As one might expect, coexpression with chaperones can reduce the formation of aggregates in some cases. Chaperones have been demonstrated to affect aggregation not only by improving recovery of soluble protein but also conversely to promote aggregation when present at high levels. When aggregates form *in vivo*, sequestration mechanisms exist that recognize aggregated species and shuttle them to designated storage locations within the cell, such as the bacterial inclusion body and aggresome or newly discovered IPOD and JUNQ sites in eukaryotic cells.\(^10,11\) When the cellular machinery is overwhelmed by excessive damage to normal proteins or by mutations that generate a less stable form of a protein that accelerates aggregation, disease or death may result. Evidence for this is found in that increased amounts of proteosomal and chaperone proteins are found colocalized with aggregates in these inclusions.

Recombinant expression has become an increasingly important method for producing large amounts of protein for therapeutic and biotechnology applications. Production of recombinant protein is often frustrated by aggregation in the host. Yield can sometimes be improved by decreasing the temperature at
which the protein is made or by coexpression with chaperone proteins (e.g., GroEL) to aid folding *in vivo* and to reduce sequestration to inclusion bodies. Nonetheless, proteins are often shuttled to inclusion bodies. Aggregated proteins, however, may be folded *in vitro* from the insoluble state. Single-domain proteins less than 150 residues, which are directed to inclusion bodies, can sometimes be extracted from the solid aggregate and refolded. Denaturing conditions are used to disrupt associations between chains, and the denatured material is diluted into a non-denaturing solution so that it may refold into its native form. When the native form of the protein contains disulfide bonds, folding is carried out under defined redox conditions to facilitate proper disulfide formation. This approach is not very efficient, typically resulting in a substantial fraction of the protein returning to an insoluble state. This observation suggests that proteins may follow different pathways during the course of folding, of which only some are productive. Very limited success has been had using this approach with large, multidomain proteins or those with numerous or more complex post translational modifications. The difficulty in refolding these proteins probably derives from increased competition between alternative interactions with those of the native state. These incorrect associations may lead to misfolding when the rate of protein production or the context in which the protein is produced is altered. Addition of chaperones at the dilution step has been used to enhance refolding of proteins that otherwise aggregate. The strategy is also being applied to stabilize purified proteins during storage. Once the active form is purified, proteins are commonly maintained at cold temperatures to restrict their conformational flexibility and to preserve their structural integrity.

Protein folding and unfolding may not follow the same pathway. Some proteins fold and unfold reversibly, yet may accomplish each event using a different approach to overcome transition state barriers. Other proteins require assistance to attain their native conformation but subsequently are quite resistant to unfolding. These observations suggest that a protein’s ability to arrive at its native conformation and to maintain it is not necessarily ruled to the same degree by the same parameters. Proteins are not static and undergo a variety of different types of conformational fluctuations. The range of states sampled is dictated by noncovalent interactions that stabilize the native fold and the effect of external influences like temperature and solution conditions on their interactions. Factors that drive folding influence the stability of the folded form, but coincident interactions that develop as a consequence of the folded state also impact retention of the native fold and help determine the frequency of transitions to partially unfolded conformations and subsequent progression to aggregated states. Once the folded protein is obtained, stabilization against transitions to aggregated states becomes a critical issue. Understanding how to prevent aggregation has primarily been based on empirical studies. The next objective is to elucidate the mechanisms that determine how transitions from the native ensemble promote aggregation.
1.1.1 Structural Features of Proteins

Proteins are linear polymers. Their primary structure is composed of 20 naturally occurring amino acids having diverse chemical properties. The amino acids are typically alpha amino acids and have L chirality. Each is joined by a peptide bond, which has a planar character that restricts the conformational freedom of the backbone of the polypeptide chain. As such, common structural features are observed among folded proteins, most broadly falling into the categories of alpha helix, beta sheet, turns, and disordered regions. These first three secondary structural elements are developed as a result of hydrogen bonding interactions that involve atoms from within the polypeptide backbone. Disordered regions lack such hydrogen bonding patterns. Alpha helices are almost always right handed and have a register in which the carbonyl oxygen from residue i forms a hydrogen bond with the amide proton four residues to its C-terminus (i + 4). The alpha helix contains 3.6 residues per turn, and due to the slight offset in vertical alignment, a secondary twist develops with elongation. Rarely, a short $3_{10}$-helix has been observed to form, in which i to i + 3 bonding occurs. The even rarer $\pi$-helix utilizes i to i + 5 bonding. Due to favored dihedral backbone angles and steric constraints, alpha is the most favorable helical organization. It often occurs in isolation, whereas the other two forms are found only in small segments in folded proteins in which the structural context provides stabilization for these less favorable structural elements. Beta-sheet structure is also favorable and can arise from a parallel or antiparallel alignment of the strands. The pattern of hydrogen bonding differs between these two sheet organizations. Antiparallel strands may form from contiguous or discontinuous primary structure, but parallel association necessarily occurs between sequences that have intervening secondary structural elements.

Contiguous stretches of repeated H-bonds stabilize each structural element and help compact the polymer within local regions of the sequence. This limits the number of possible arrangements between distal segments, which also facilitates the establishment of a preferred three-dimensional conformation (tertiary structure). Three-dimensional coalescence into a compact state generally relies on interactions between amino acid side chains. The diverse chemical composition of the side chains produces both attractive and repulsive forces, and the native configuration derives from the formation of the most energetically favorable associations between distal moieties that stabilize the packed arrangement within a folded domain. The majority of interactions that contribute to protein folding are noncovalent, but covalent bonding between the thiol-containing moieties of cysteine residues may occur to generate a disulfide bond under oxidizing conditions. Disulfide bonds are common among secreted proteins, where they often greatly enhance the ability of the protein to resist unfolding.

Globular domain folds are classified into families that range from all helical to mixed alpha-beta to all beta composition. Regardless of the domain archi-


ture, separate polypeptide chains can further associate into homotypic or heterotypic oligomers to yield a quaternary structure. Individual subunits in an oligomeric complex can simply physically associate based on surface complementation, but they may also be covalently tethered. Covalent attachment ensures close proximity and is most often accomplished through intermolecular disulfide bonds.

Noncovalent association between subunits vary in affinity based on the same principles that dictate protein folding, and several modes of interaction have been described, including lock and key, induced fit, and preexisting equilibrium/conformational selection mechanisms. Lock-and-key binding implies the structure is unaltered by the binding event. Induced fit models suggest that the protein adopts a new state in response to binding to its partner, whereas conformational selection indicates that in the associated complex, an existing state is stabilized. Analogous modes of interaction may also apply to protein–protein associations that pertain to physical aggregation. A relatively new area of investigation has demonstrated the diversity of conformations that can result from the same sequence. For example, natively disordered proteins have been suggested to adopt distinct conformations in different contexts to perform discrete functions. Recent studies also reveal that globular proteins can maintain more than one unique stable conformation. Moreover, crystal structures are often reported for the same protein in distinct oligomeric states depending on solution conditions. These findings suggest that preferences in the conformation of a protein, even those below the current limit of detection, are influenced by the context in which the protein resides.

1.1.2 Structural Features of Protein Aggregates

Macroscopic attributes of aggregates have been described from data acquired using a variety of microscopy techniques. In the most general terms, the morphological features commonly observed are usually categorized as amorphous or fibrillar. Amorphous aggregates are present in inclusion bodies in vivo and often emerge during the course of processing and storing protein samples. Amorphous aggregates lack long-range order and are often opaque if they are not soluble. They were originally thought to contain completely unfolded material held together by random associations between hydrophobic residues. The hypothesis that amorphous aggregates lack discernable structure was derived from (1) early observation that harsh denaturants like sodium dodecyl sulfate (SDS), urea, and guanidine-HCl (Gdn) are required to resolubilize proteins from inclusion bodies; and (2) a lack of data concerning structural features owing to the fact that amorphous materials often scatter light, interfering with the spectroscopic analyses typically used to characterize structure. In contrast, some aggregates retain native activity, and the active form of some proteins can only be recovered from inclusion bodies when mild denaturing conditions are used, whereas aggressive denaturation
AGGREGATION OF THERAPEUTIC PROTEINS

leads to an inability to refold the protein (e.g., hGH), suggesting that the species present in the inclusion body contain elements of native structure that facilitate refolding. Additionally, staining methods have been used to reveal the presence of regular structure within amorphous aggregates. Staining of some aggregates and not others by dyes like Congo red (CR) and thioflavin T (ThT) suggests that proteins directed to inclusion bodies contain at least some ordered structural elements common to fibrillar aggregates. Many aggregates have increased beta-sheet content and diminished alpha helicity compared to the native state, which is presumably developed through intermolecular contacts. Although the molecular organization of amorphous aggregates remains rather coarsely described overall, the Weliky lab recently provided evidence for native-like structure within amorphous aggregates. Their examination of influenza virus hemagglutinin within an inclusion body using solid-state nuclear magnetic resonance (NMR) indicates the presence of a substantial retention of helical structure. The residues that form helices in the aggregate correspond well to those present in the native conformation observed in the available crystal structure. This result, combined with the fact that amorphous aggregates from inclusion bodies bind CR and ThT, suggests that aggregates are composed of a combination of native- and fibrillar-like structures.

Fibrillar aggregates are commonly, but not always, associated with the formation of amyloid plaques and are named for their long, thin, fibrous shape. Great diversity in the diameter, length, and interconnection between amyloid strands has been observed, but each have hallmark characteristics of being birefringent when stained with dyes such as CR and of being detergent insoluble. CR and ThT bind to fibrils and were first used to identify the presence of a regular structure in fibrillar aggregates derived from proteins of diverse composition. Fibril variation depends on the protein from which they are derived and the conditions under which they are made. Such macroscopic differences suggest that despite having common features, structural diversity exists. Fibrillar aggregates contain long-range order and, as a result, have been described in greater detail than amorphous forms. Although fibrils are not amenable to atomic-level structure determination by X-ray crystallography or by solution NMR, analysis using medium to high-resolution techniques and reconstructive methodologies has begun to provide molecular details of their organization. These aggregates have been shown to be composed of substantial amounts of beta-sheet structure and lacking in helical content. When helical structure is present, it is significantly diminished compared to the native state of the corresponding protein. Both parallel and antiparallel sheet orientations have been observed in fibrils using IR spectroscopy, electron microscopy, and NMR. Aggregates composed of antiparallel organization are typically formed from short peptide strands. Longer fragments in which the same sequence is embedded arrange themselves preferentially into parallel strands, as was observed from studying Aβ peptides. Most fibrils are composed of parallel strands.
One particularly interesting feature of fibrillar aggregates is that they seem to depend largely on backbone interactions that are distinguishable from those of folded proteins. Antibodies have been generated that recognize only the fibrillar form of a peptide and not the monomer or other intermediate aggregation states.\textsuperscript{22,23} Because these antibodies can also bind to fibrils composed of entirely unrelated peptide sequences, it was concluded that the antibodies must recognize a common backbone configuration that exists uniquely in fibrillar structures. Higher-resolution X-ray diffraction and NMR data lend credibility to this assertion by revealing that amyloid fibrils have a cross-$\beta$ structural organization (Fig. 1.1) that has not been observed in globular proteins. The cross-$\beta$ spine has been shown to occur through a parallel arrangement of beta strands\textsuperscript{24} in which the side chains between two facing sheets interdigitate, forming a steric zipper (Fig. 1.2).\textsuperscript{25-27} The zipper interface is tightly packed and is completely dehydrated.\textsuperscript{26} The strands run perpendicular to the long axis of the fibril with 4.7- to 4.8-Å spacing between each strand.\textsuperscript{28} The cross-$\beta$ organization has been observed in aggregates of several unrelated amyloid peptides, suggesting this is a regular structural element in fibrillar aggregates. The same peptide sequence, however, has the ability to participate in different morphologies, and these are hypothesized to derive from unique packing arrangements of the spine. The Eisenberg lab proposed a structure for the fibril that is consistent with distances established by X-ray diffraction, electron paramagnetic resonance (EPR), and intrinsic fluorescence of multiple peptides.\textsuperscript{26} Viewed down the long axis, it is a dimer of dimers with a left-handed twist of 3.4° per layer, which has a diameter of 64 Å when based on a 20-residue peptide (Fig. 1.3). Earlier models are similar in many regards. The serpentine model of the amylin fiber is analogous but is based on fewer sheets.\textsuperscript{29} The Tycko model has four sheets but does not have as tightly interdigitated side-chain packing because it was based on distance restraints derived from solid-state NMR experiments that indicate closer interactions between certain side chains than are used in the Eisenberg model.\textsuperscript{27} Recalculation of the Eisenberg
model to include the NMR-based side-chain interactions resulted in poorer alignment between the sheets. The common element of the cross-β arrangement among unrelated sequences indicates that a backbone conformation amenable to beta structure and compatible interaction between side chains in the strands stabilize fibrillar aggregates. The disparities between the models suggest that differences in sequence, which affect side-chain packing, may be responsible for the morphological differences observed in aggregates.

The phenomenon of aggregation is complex and seems to involve a series of intermediate structural states, some of which may be amorphous, along the path toward forming highly ordered fibrillar amyloid structures (Fig. 1.4). Some of these states have been described in varying degrees of detail. Many plaque-forming proteins undergo an early transition from native monomers to oligomeric aggregates before fibrils are detected. The shape of many such species appears spherical or ellipsoidal (e.g., Aβ, α-synuclein, prion protein,
cystatin C). Their diameters vary, but often it has been shown that a discrete number of protein molecules are involved, suggesting that specific structural contacts are made between each polypeptide chain. These contacts are often so stable that they resist dissociation in the presence of SDS, earning the

Figure 1.3. Model of macroscopic fibril structure composed of four sheets shown looking down the long axis (A) and from the side, showing the twist (B). From Reference 26.

Figure 1.4. Peptide fragments arrange in beta barrel formations during simulations. Closed or open barrels result from different peptide sequences. Reproduced with permission from Reference 33.
AGGREGATION OF THERAPEUTIC PROTEINS

designation stable aggregates. Aβ forms several distinct stable oligomers depending on a variety of factors. In the presence of trimethylamine N-oxide (TMAO), Aβ (1–40) generates elliptical aggregates of 5.0 × 1.5 × 6.0 nm, which must be composed of four or five strands. SDS-insoluble Aβ oligomers composed of 3–24 strands have been observed, and metastable tetramers and dodecamers have been shown to convert into fibrillar aggregates. It is evident in some micrographs of protein aggregates that fibrils emerge from association between oligomers, as beads on a string-like structures are apparent in the images preceding fibrillation. Because fibrillar forms are easily detected, it was originally thought that the various protein solubility diseases were actually directly caused by the fibers. More recent evidence indicates that fibers may be more a consequence of disease processes, which are instead caused by oligomeric aggregates. In support of this, Aβ species having globular structures of 4–5 nm have been shown specifically to be toxic, whereas smaller oligomers and larger fibrils are less damaging. Experiments involving other amyloidogenic proteins, including transthyretin, islet amyloid protein, α-synuclein, immunoglobulin light chain, and β2-microglobulin (β2m), as well as non-disease-related proteins, reconfirm that in general, non-fibrillar, SDS-stable, low-molecular-weight aggregates are cytotoxic. This common finding among so many unrelated sequences is suggestive that similar mechanisms and structures underlie protein pathogenesis. Although the exact structure of toxic species is unknown, these soluble aggregates all possess the ability to interact with and cross membranes and to disrupt ionic gradients. As such, it was proposed that beta barrels might form in the membrane, mimicking the structure of ion channels (Fig. 1.4). Amyloids have been induced to form in the presence of membranes, detergents, surfactants, polyanions, and compounds that change the dielectric of the solvent, which suggests that conformational changes mediate aggregation. Computational analyses of peptide aggregation lend support to the formation of β-barrels by aggregation-prone sequences. Depending on the chain length and number of strands that participate in oligomerization, either open or closed barrels may form. Closed barrels convert to orthogonal sheets, while open barrels generate parallel sheets. This finding suggests that different proteins may undergo semi-common intermediate aggregation states that lead to distinctly different final morphologies. Pathogenesis may not be a consequence of a single universal mechanism since linear protofibrils greater than 400 nm composed of tau protein are notably toxic as well. Higher-resolution structural information about each of these systems is needed to investigate the β-barrel hypothesis and to better understand the relationship between aggregate structure and membrane permeability.

Most proteins unfold as the temperature is elevated, and at temperatures below their unfolding transition (Tm), where only partial unfolding is observed, aggregates begin to form. Typically, the aggregates are amorphous, as detected by light scattering and/or by the formation of opaque precipitates. For example, water-soluble aggregates of ovalbumin were induced to form by heat denatur-
ation, and these soluble aggregates were observed to develop cross-β structure as recognized by ThT binding. The finding suggests that the formation of amorphous aggregates may involve mechanisms common to fibrillar aggregates, but no high-resolution structural information is available to validate this inference. Because it is very difficult to ascertain atomic-level structural details from amorphous systems, identification of parallels between amorphous aggregates and those having a long-range order provides a basis for a better understanding how amorphous aggregates are stabilized.

### 1.2 PARALLELS TO PROTEIN FOLDING

Anfinsen’s Nobel prize-winning studies on ribonuclease demonstrated the seminal importance of hydrophobic collapse in protein folding, generating the *thermodynamic hypothesis*. This hypothesis states that the condensed form of the polypeptide is spontaneously achieved in aqueous solution because burial of hydrophobic moieties in the protein core is energetically favorable. Numerous experiments have been conducted on many proteins since then, reinforcing the general nature of this conclusion for proteins. For practical reasons, in the case of protein folding, it has been presumed that the free energy thermodynamic minimum is met in the native state; nevertheless, conversion from the native conformation to more stable aggregated states regularly occurs from the normal range of fluctuations experienced by proteins. The factors that drive proteins to the condensed state are determined by thermodynamics and kinetics. The native state is defined by the thermodynamic minimum of the active protein under physiological conditions. Nonnative states that are more thermodynamically stable may emerge, but they are not active and/or kinetically accessible on the biologically relevant time frame (ms-sec) of protein folding.

Although the thermodynamics of protein folding were clearly established by Anfinsen and colleagues, the mechanism or pathway for folding was not. Originally, protein folding was presumed to occur via a random process in which the polypeptide chain could adopt any possible conformation on the way to the native state. In this model, an astronomical number of conformations are possible and equally probable. Levinthal pointed out the paradox that the timescale on which proteins must fold is many, many orders of magnitude shorter than a completely unbiased random process would require. Since experimental results show that proteins achieve native folds in ms-sec rather than the predicted expanses of years, the hypothesis emerged that folding must be directed in some way toward the native state. Since then, many have worked to identify and explain the parameters that permit rapid folding of proteins to their native state. Proteins having different amino acid sequences have different rates of folding and unfolding as well as differing stabilities. Single amino acid substitutions in a protein can dramatically affect these processes. These observations provided the foundation for the idea that the amino
acid sequence encodes structural information that influences folding and stability. As such, some have sought to identify sequence motifs responsible for aggregation, but no well-defined common sequences have been found. Instead, it appears that general features of the polypeptide sequence direct cooperative folding. The properties of the side chain influence both the local backbone conformation and packing interactions, imposing restrictions on the folding pathway.

Both protein sequence and environmental conditions have been shown to affect the structural stability and aggregation of proteins. The Gibbs free energy depends on the totality of interactions within a system, and as such, the thermodynamic minimum for a protein is determined not only by the amino acid sequence but also by the environment in which it is contained. Temperature, pH, ionic strength, viscosity, counterions, and other factors have been shown experimentally to influence protein stability, indicating that environmental conditions are critical determinants of the lowest energy conformation of a protein. In many cases where misfolded proteins cause disease, mutation is responsible for destabilizing the native state under normal conditions, thereby lowering the barrier to structural transitions that generate non-native conformations and promote aggregation.

1.3 VIEWS OF PROTEIN STABILITY AND AGGREGATION

Views of protein stability have largely emerged from studies of well-behaved, single-domain proteins that undergo reversible folding to unfolding transitions. Numerous experiments have been carried out that employ thermal or chemical denaturation to examine folding or unfolding by following fluorescence or absorbance changes, hydrogen/deuterium exchange, or structural transitions by circular dichroism (CD) or NMR. Reversible analyses are almost exclusively performed using chemical denaturants, because thermally induced unfolding typically leads to rapid aggregation as the structure begins to be altered. At high concentrations, chemical denaturants compete for binding to the moieties involved in self-association, thereby promoting retention of the unfolded state. The most frequently characterized and well-described systems display two-state kinetics, although a few examples of multistate kinetics have been reported. These simple systems have obvious utility for investigating fundamental influences on protein folding, and the findings are relevant to protein aggregation. The fundamental parameters that influence the behavior of proteins in solution and the models that have been developed to explain their behavior are discussed below.

1.3.1 Physicochemical Properties of Proteins

Although proteins appear complex, their behavior is ultimately dictated by basic chemical and physical principles. From a thermodynamic perspective,
collapse of the polypeptide chain is an organizing event and also requires that
water be removed from parts of the protein that form the core of the folded
species. The energetic cost of removing water and the entropic penalty (\(\Delta S\))
must be compensated in order for folding to occur. This is generally achieved
by a decrease in the enthalpy (\(\Delta H\)) of the protein. The free energy (\(\Delta G\)) in
the native folded state is typically 5–20 kcal/mol more stable than the unfolded
state.\(^{35,40}\) This \(\Delta G\) value represents only a fractional difference in the total
energy of both the folded and unfolded state.\(^{40}\) It is approximately equivalent
to two to four hydrogen bonds, and even small proteins employ tens of intra-
molecular H-bonds in their fold,\(^7\) thus indicating that the sum of many small
energy differences contributes to achieving a folded state. The major factors
that have been shown to play a role in the specificity and stability of protein
folds are derived from the chemical attributes of the amino acid sequence,
including hydrophobicity, aromatic stacking, electrostatic interactions, steric
constraints, and hydrogen bonding (secondary structure), as well as the effi-
ciency of packing in the core and surface properties of the folded molecule.
The contributions of each component to protein folding and stability are dis-
cussed below.

\[\Delta G = \Delta H - T\Delta S.\] (1.1)

Hydrophobicity and apolarity have been shown to be the primary factors
that drive proteins to the condensed state because of the large heat capacity
change associated with this process. The driving force for hydrophobic
collapse is modulated by the temperature, pH, and dielectric constant (\(\varepsilon\)) of
the solvent. At extremes of any condition, structural and dynamic changes
are apparent. Low pH (<3) conditions often cause native folds to relax, trans-
forming into molten globule (MG) states, especially at high ionic strengths.\(^{42}\)
The MG retains much of the secondary structure of the native state, but ter-
tiary packing is disrupted such that the residues in the core are solvated effi-
ciently. This has been shown by H/D exchange experiments and by heat
capacity measurements, which indicate these properties are more akin to the
unfolded state than the native form.\(^{43}\) Titration analyses performed on numer-
ous proteins reveal that the stability of the fold, as determined by changes in
structure and aggregation, is also significantly affected by pH.\(^{44}\) In this regard,
it has been shown that benzyl alcohol, a hydrophobic additive used as an
antimicrobial agent, destabilizes the tertiary structure of recombinant human
interleukin-1 receptor antagonist (rhIL-1ra) and granulocyte colony-stimulat-
ing factor (GCSF), inducing aggregation.\(^{45,46}\) Changes in solution conditions
that diminish the driving force for hydrophobic association concomitantly
interfere with core packing and disrupt the native fold. Increasing temperature
causes a decrease in dielectric constant and, above approximately room tem-
perature, an increase in the solubility of hydrophobic groups.\(^{47}\) Because the
difference in free energy between the folded and unfolded state is relatively
small, changes in temperature and the dielectric properties of the solution
presumably lead to unfolding through alteration of the driving force for hydrophobic collapse. The dielectric constant for water at 20°C is approximately 80, while it is estimated that the ε value of the protein core is 5–20. The dielectric of water is strongly affected by temperature such that its ε decreases with increasing temperature. The dielectric of hydrophobic moieties are largely unaffected by increasing temperature in the absence of water, but hydration due to increased thermal motion increases the dielectric (Fig. 1.5). As such, raising the temperature lowers the barrier to unfolding, which is consistent with experimental observations.

Aromatic residues (Phe, Trp, and Tyr) contribute to hydrophobic interactions that stabilize the core because they have relatively large apolar surface areas. Additionally, they provide enhanced stability when arranged in a stacked configuration, where they not only maximize van der Waals contacts but benefit from pi–pi interactions. Moreover, the stacked orientation is an efficient packing arrangement. A well-packed core further contributes to protein stability. Aromatic stacking has been proposed to stabilize amyloid fibrils, and a parameter-free model indicates pi-stacking should facilitate aggregation. Pi bonds produce a dipole, which also enables favorable interaction with cations. There are many examples of structures in which Lys and Arg side chains are found to interact with aromatic rings in folded proteins. These associations are believed to be primarily electrostatic, but regular distances and geometries also suggest hydrogen bonding may be a mode of

**Figure 1.5.** Graph showing the effect of temperature on the dielectric constants of water, methanol, and ethanol. The dielectric constants of the alcohols are virtually unchanged with increased temperature, whereas the value for water is dramatically altered. From http://www.greenfluids.org/SubcriticalWater.htm (accessed January 2009).
interaction. Aromatic–aromatic and aromatic–amine interaction also provide an enthalpic contribution to protein stability.

Charged residues, including Glu and Asp (acidic) and Lys and Arg (basic) most often appear on the surface of folded proteins. Because they interact favorably with water, their solvent accessibility is a dominant factor in the overall solubility of proteins. The isoelectric point (pI) of a protein to some extent determines the solubility of a protein at a given pH, with the lowest solubility theoretically occurring at the pH equivalent to the pI. Titration of the pH away from the pI to either more basic or more acidic conditions often improves solubility within the limit of structure retention. Although His is a basic amino acid, it commonly exists in the neutral state and is located in the core of folded proteins. The pKa of its imidazole is 6.7 in solution but can be perturbed substantially in the context of a folded protein due to the local microenvironment. His participates in hydrogen bonding, where it can act as a donor or as an acceptor. It is paired with buried Tyr in many enzymes, where it lends specificity to the interaction between two structural regions that undergo a conformational change induced by substrate binding and turnover. The uncharged species is easily accommodated in the hydrophobic core because it can add stability through hydrophobic and aromatic interactions as well. Electrostatic interactions involving Asp, Glu, Lys, and Arg are also observed in the core of folded proteins in the form of salt bridges. The pairing of a positive and negative charge permits the incorporation of these residues into the apolar core and provides structural specificity to the packing arrangement. Uncompensated charge, however, disrupts hydrophobic packing, decreasing stability. Mutations that result in protein aggregation often introduce uncompensated charge or polar moieties into an otherwise hydrophobic environment, thereby inducing partial unfolding in which significant amounts of apolar surface area become solvent accessible. Interestingly, evolution seems to have selected for protection against aggregation at the edge strands of beta sheets using this strategy. Lys is often positioned in the center of terminal strands where it disfavors interaction between the outside edges of β-strands. This idea has been applied to test the aggregation potential of amyloid peptides. The substitution of a hydrophobic with a charged residue ameliorates aggregation and, when added at high concentrations, promotes dissociation of aggregated species. Modeling studies predict that an amyloid peptide is bound by the mutant peptide and this disrupts the pattern required to support fibrillation.

The most important steric constraint that limits the number of possible backbone conformations is the planarity of the peptide bond. The original proposed models for possible secondary structures did not account for this chemistry and consequently presented many more configurations than are possible. It was the subsequent recognition of the significance of the planar bond in protein structures that earned Linus Pauling a Nobel Prize. The consequence of this constraint is obvious in the now standard Ramachandran plots of globular proteins where $\phi/\psi$ dihedral angles largely cluster into
two main regions corresponding to the $\alpha$-helix and $\beta$-strand conformations.\textsuperscript{68} Quite a range of deviation is observed around the optimal angles for alpha helix and beta structures, indicating substantial flexibility is permitted in the formation of these regular elements. Secondarily, dihedral angles are affected by steric restrictions that the side chain imposes on the backbone. The degree of solvation and preferred dihedral angles of the backbone are correlated with the degree of solvation of the side chain, such that amino acids and peptide sequences have an inherent propensity for specific secondary structures.\textsuperscript{69–71} The degree to which this bias exists depends on the amino acid and on the sequence in which it is embedded. Polyalanine sequences have a strong tendency to form alpha helices. Leu is found approximately 60% of the time in helices and 25% of the time in beta strands, whereas beta-branched amino acids (Val, Ile) tend to adopt an extended beta structure, such that Ile is located in beta strands in about 75% of sequences. Many sequences have only a marginal bias such that the actual development of a secondary structure depends on context, which may be derived from neighboring sequences and structures in the protein or may be influenced by solution conditions. There are several pieces of evidence that support this claim. Peptides that form stable helices or strands in folded proteins often exist as random coils in isolation.\textsuperscript{72,73} Alterations in the dielectric constant of the solvent, commonly by addition of alcohols such as TMAO or trifluoroethanol (TFE), induce secondary structure in random coil peptides.\textsuperscript{46,74} At low concentrations of these alcohols, alpha helicity increases, but at concentrations where the cosolvent comprises a significant fraction of the volume, beta structure has been observed. It is important to note, however, that peptides from specific regions in proteins do not necessarily convert into the same structures in the presence of alcohols.\textsuperscript{75} These additives change the dielectric of the solvent such that hydration of the backbone is altered by the differences in backbone and side-chain solubility compared to water.\textsuperscript{70}

All secondary structures are stabilized by hydrogen bonds.\textsuperscript{7} Alpha helices emerge from local, sequential interactions, with their bonding pattern involving repeating interactions between each CO at residue $i$ and NH at residue $i + 4$. Sequences that contain hydrophobic residues every seventh and third or fourth position tend to form helices because the pattern of apolar side chains is in register with the 3.6 amino acids per turn inherent in the $\alpha$-helical structure. This produces an amphipathic arrangement and supports stabilization of the helix by interaction between its hydrophobic face and another hydrophobic surface. Often the partner is another helix, as seen in coiled-coil structures or helical bundles, but interaction with beta sheets is common as well. When polar or charged residues occupy the remaining positions, the opposite face of the helix is amenable to interaction with water resulting in a soluble arrangement. Helices composed mostly of hydrophobic residues are poorly solvated and must be embedded in the core of a larger assembly. Beta sheets rely on the formation of hydrogen bonds between residues that are distant in the sequence. They have been shown to form from sequences in
which hydrophobic (H) residues alternate with polar (P) residues, that is, HPHPHP. This pattern generates compatible interactions between strands and produces the expanded hydrophobic surface necessary to sustain collapse as well as an apposed hydrophilic surface to support the solubility of the folded form. Consistent with this, a large number of proteins fold into beta sandwiches in which their hydrophobic surfaces face inward to stabilize the interaction between the two sheets composing the sandwich, and their outer surfaces are amenable to solvation by water. In the cross-beta structure formed in amyloids, the hydrophobic residues are positioned on the interior of the strand and the orientation of the strands is straight with respect to one another, such that the addition of strands to both ends is favorable. Aggregation involves not only hydrophobic association but also depends strongly on backbone hydrogen bonding to form an intermolecular beta structure. Methylation of the backbone amide nitrogen inhibits aggregation of amyloid peptides despite the presence of an extensive hydrophobic surface area and strongly indicates that intermolecular hydrogen bonding is essential to the formation of stable aggregates.

The propensity of individual amino acids to form alpha and beta structures has been characterized in a variety of ways. Statistical approaches have been used to assess the frequency of each amino acid in helices and sheets in known structures and in the NMR chemical shifts of backbone resonances. Experiments were performed on helical peptides to characterize the helical propensity of each amino acid. Theoretical studies revealed that the basis for the observed helicity has to do with the dihedral angles of the side chain, because entropic losses are incurred upon folding. Alanine most strongly favors the helical conformation because no loss occurs. Modeling of dipeptides has been performed to obtain information on preferred backbone torsion angles. The dominant contribution to beta-sheet propensity was found to be the avoidance of steric clashes between the backbone and the associated side chain. There are contextual constraints imposed on beta-strand association, but this study indicates that inherent beta-propensity is strongly tied to local features. An interesting outcome of the study was that Asn displayed an exceptionally high propensity for beta-structure, which was attributed to the unique hydrogen-bonding capabilities of its side chain.

An important additional consideration is the kinetics of folding as it pertains to individual secondary structure elements. The contacts involved in beta structure are more disbursed and cannot assemble as rapidly as adjacent interactions within individual helices because the segments forming beta-strands are discontinuous. This is not to say that all helices form equally rapidly, as was noted above in the discussion of sequence effects. The antiparallel arrangement may form more rapidly when a short turn, particularly a well-defined turn, separates the strands, whereas parallel configurations are necessarily separated by long stretches of residues not involved in the sheet. Proteins with such arrangements often receive assistance from chaperones to fold in vivo.
sequences are separated by helical segments. Refolding experiments show that rhodanese aggregates readily following dilution from denaturants, despite a significant retention of the secondary structure. Inclusion of several chaperones is necessary to prevent aggregation and facilitate its refolding. The results of this study indicate that both extended structure and an MG-like state exist, which must be stabilized during the folding process to avoid aggregation. The inference then is that the sequences involved in sheet formation are protected from aggregation by chaperones until they are oriented properly with respect to each other. Antiparallel sheets composed of strands with long intervening sequences are akin to parallel configurations in this regard, indicating a longer time may be needed for their formation. This difference in kinetics may provide an explanation for why alpha to beta conversion is observed in aggregation. When alpha helices are destabilized, beta strands that exist may become amenable to rapid intermolecular association since they are no longer protected by intramolecular structural elements. A survey of protein structures revealed that beta sheets are often protected from intermolecular association in their native conformation because they are covered by helices. Alternatively, the newly uncoiled region may sample β-conformations more frequently and may directly participate in aggregation. If a competition exists between the two secondary structures in the same sequence, structural stabilization and conditions that favor helical organization would prevent aggregation. If the exposure of the existing beta structure is responsible for aggregation, then preservation of helicity in the absence of tertiary structure retention would be insufficient to prevent aggregation.

Although most aggregates have diminished helical content and/or increased beta structure, it is possible that association between helices could confer nonnative oligomerization via open domain-swapped conformations, which subsequently may promote further aggregation. Even when an equivalent amount of hydrophobic surface area is buried, helices lack the hydrogen bonds that stabilize intermolecular beta interactions. The upshot is that intermolecular beta sheets are more likely to produce more stable (irreversible) aggregates than alpha helical associations, which is consistent with available data.

**1.3.2 Surface Properties and Packing Arrangements**

The driving force for folding is primarily burial of apolar surface area, but the composition of the outside surface also has significant implications for stability once the native fold is attained. Polar and charged residues constitute the majority of the surface area on protein exteriors because they are more easily solvated by water than by hydrophobic side chains. On average, greater than 70% of the surface area of monomeric proteins is hydrophilic. Charged moieties are more easily hydrated than non-charged species and have a greater driving force for interaction with water. The number, density, and location of charged residues on the surface determine solubility and also exert a profound
influence on the stability of proteins. In globular proteins, the more charge per square angstrom, the greater the thermodynamic stabilizing force on the protein. There seems to be an optimal charge density, however, because proteins having extremely high percentages of charged residues are often natively disordered. Thermophilic organisms are adapted to function at greatly elevated temperatures, and their proteins can occasionally resist thermal unfolding above 100°C. A computational study comparing the stability of thermophilic and mesophilic proteins showed that the enhanced stability of the former was imparted by surface residues. The hydrophobic core of the proteins contributed equally to stability, but the increased density of packing among the surface residues of the thermophilic proteins correlated strongly with stabilization. Proteins from thermophiles were noted to display higher percentages of Lys, Arg, and Glu and lower fractions of Ala, Asp, Asn, Gln, Thr, Ser, and His on their solvent-accessible surface than mesophiles. Additionally, increased enthalpy changes at the melting temperature have been observed for proteins with higher thermostability, suggesting that electrostatic moieties participate in preventing unfolding. This principle may also underlie low pH induction of the MG state. By protonating the acidic groups and by reducing the net surface charge, the differential solubility between surface and core residues is decreased, thus lowering the barrier to unfolding. Not only is the total number of charges important, but the density and location of charged moieties also influence stability. Because like charges repel each other, compensation is important for stability, such that stabilization is bolstered by attractive interactions outweighing repulsive forces. Individual salt bridges on the surface have been shown to contribute significantly to the stabilization of many proteins. Networks of charged residues and hydrogen bonds are found on the surface of proteins and have been shown to enhance stability synergistically. Salt bridge formation on the surface of a protein can improve structural stability by contributing a favorable tertiary interaction between nonsequential residues. Clustering of positive or negative charges on compatible surfaces can also result in intermolecular ion pairing to generate soluble or insoluble oligomers or fibrous structures. Oligomers that retain their native structure can be solubilized reversibly by pH adjustment or high salt, as has been shown for the CC chemokines. If the interaction alters the tertiary structure, aggregation may be irreversible as a result of more extensive structural rearrangement.

Although most of the protein core is composed of hydrophobic moieties, structures in the Protein Data Bank (PDB) show that acidic and basic side chains are often buried in the protein core. The proteins exist in stable folded structures as long as the charge is offset in a way that its presence is more stabilizing than destabilizing. This can be accomplished by the pairing of oppositely charged groups. Salt bridges, N–O pairings, and longer-range ion-pairings help decrease the energy cost associated with burial in a nonpolar environment. The pairing is most effective when the center of charge and at least one N and O atom from each residue are within 4 Å, constituting a salt
bridge. The energy is higher when only the second criterion is met but is still effectively stabilizing in many proteins. Longer-range pairings are stabilizing when a larger network of interactions, often involving hydrogen bonds, is present or alternative pairings exist within ~5 Å. An intermolecular salt bridge has been proposed to help stabilize fibrillar aggregates of Aβ, in which Lys16 interacts with Glu22 in the antiparallel configuration.\textsuperscript{21,93} Several modeling studies confirm charge complementarity and support interstrand and beta-sheet association.\textsuperscript{94–96}

An alternative means of accommodation of charged moieties in protein cores is perturbation of side-chain pKa or stabilization through hydrogen bonding.\textsuperscript{97} The pKa of the hydrated carboxylic acid moieties in Asp and Glu in solution is 3.5 and 4.1, respectively.\textsuperscript{98} The presence of adjacent groups greatly affects the pKa of these residues, causing it to range from 0.5 to 8.8. For example, the decreased dielectric constant experienced by a moiety within the hydrophobic protein core can raise pKa values sufficiently to make the carboxyl group neutral at physiological pH. pKas are much more substantially altered by hydrogen bonding, which lowers the pKa. Each H-bond depresses the pKa by approximately 1.6 and contributes 2 kcal/mol toward stabilizing proteins.\textsuperscript{97} Asp76 in RNase T1 makes three strong hydrogen bonds, and its pKa is 0.7. Disruption of such hydrogen bonds in RNase T1 and in many other proteins reduces their thermal stability.\textsuperscript{97,99} Groupings of charged residues on the surface of a protein can also influence the net charge by perturbing the pKa of residues in close proximity, but this usually has only a small effect, inducing fractional changes in the measured pKa. Structures of Aβ fibrils align such that strings of Glu residues are positioned adjacent to each other along the long axis of the fiber. The Glu side chains are surrounded by a high density of hydrophobic moieties on both sides. The high degree of hydrophobicity and close proximity of like charges both should contribute to elevating the pKa of the carboxylate group, lowering the barrier to association between sheets.

It follows from earlier arguments that the amount of apolar surface area exposed on the outside of folded proteins should influence their solubility. One way that surface hydrophobicity is mitigated to preserve solubility is through protein oligomerization. Many proteins function as dimers, trimers, tetramers, or higher-order oligomers \textit{in vivo}. Analysis of their structures reveals that the interfaces between subunits have an extensive hydrophobic contact area to support self-association. Heterodimerization also helps preserve the integrity of protein monomers with an exposed hydrophobic surface area, and isolation of one protein in the absence of its partner often leads to aggregation.\textsuperscript{100} Protein–protein interaction involves the association of complementary surfaces, which are often approximately planar.\textsuperscript{101,102} This is common among dimers that interact by abutting edge beta strands to form an extended sheet structure (e.g., λ-cro, chemokines).\textsuperscript{92,103} Association also results from interactions between hydrophobic clusters found in loops or helices (e.g., hemoglobin, PRL-1).\textsuperscript{104,105}
1.3.3 Solvent Interactions

It is apparent from countless studies that the solution environment surrounding a protein strongly influences its structural retention. The long-standing, dominant theory that explains this phenomenon argues that interactions between the solvent/cosolvents and protein are very weak and nonspecific, such that simple parameters such as osmotic pressure and surface tension dictate stabilization. The interaction between proteins and water is still poorly understood, and little is certain at this point about how stabilization is achieved at the molecular level. In aqueous solutions containing proteins, water molecules typically possess two diffusion rates: one associated with the bulk solution and another with the hydration sphere associated with the protein. Water in the hydration sphere is more static than in bulk solution, forming a layer approximately 7 Å thick around the surface of the protein. The individual water molecules can be retained from tens of picoseconds to nanoseconds, differentiating them from bulk solutions, which have residence times of approximately 1–8 ps. \(^{106,107}\) Based on residence time, little associated water is detected in unfolded proteins. This data suggest that the water molecules may be organized into a more structured network around the protein. This hydration shell is thought to promote stability in part because the water binds more tightly to the folded protein than does bulk water to the unfolded form, a simple mass action effect. The basis for increased residence time has not been quantitatively established as of yet and does not appear to correlate with solubility. Although proteins with larger charged surface areas are often more soluble and it might be expected that water would be retained around charged moieties, it has already been noted that too high a charge density often produces disordered proteins, which have reduced retention of water at their surface. Because the arrangement of water around the folded protein is rather poorly understood at the molecular level, it is possible that the shape, specific interactions that occur between moieties in close proximity on the surface of the protein, and the local structure and chemical properties of the surface-exposed groups may be major factors that contribute to the transient organization of water molecules in the hydration sphere.

Ions are also retained at the surface of proteins when present at low concentration due to field potentials induced by charged residues. \(^{108}\) The interaction of salt ions with charged residues on a protein’s surface has been modeled, and they have been found to interact preferentially based on residence time. Attraction of ions is stabilizing to a limited extent, depending on the protein. For example, lysozyme binds a Cl\(^{-}\) ion that slows thermal denaturation. High concentrations of ions are destabilizing because they outcompete the protein in binding to water molecules, often inducing precipitation through the well-known salting out phenomenon. \(^{109}\) This phenomenon may alternatively induce conformational rearrangement by shielding stabilizing interactions between charged species on the surface of the protein. Alternatively, very high concentrations of some larger ions such as amino acids can stabilize
proteins through a preferential hydration effect.\textsuperscript{110} Charge neutralization may also be accomplished through the interaction of polyanions with binding partners. Destabilization of protein structure has been reported in the presence of sulfated glycosaminoglycans (GAGs) and proteoglycans and has been found to accelerate nucleation and fibril polymerization of A\textsubscript{β}, albeit through an unknown mechanism.\textsuperscript{111} Conversely, many native proteins are dramatically stabilized by the presence of polyanionic compounds,\textsuperscript{44,112–115} suggesting that specific knowledge of the mechanisms by which each event occurs is needed to explain the differing results. One possible mechanism of aggregation is that binding may induce and stabilize conformational changes that expose the apolar surface area. Further rearrangement may then be permitted due to neutralization of charged moieties through electrostatic interactions.\textsuperscript{63} When binding leads to stabilization of a protein, the conformation would be such that the accessible hydrophobic surface area is limited or it persists in a conformation that is not amenable to self-association.

\textbf{1.4 MODELS OF AGGREGATION}

Several mechanisms have been proposed to describe the formation of aggregates that are based on the observed kinetics of assembly. Each begins with a partial unfolding of the native state. Partial unfolding permits the association of monomers through exposure of previously inaccessible residues. Metastable monomeric intermediates are thought to be responsible for aggregate formation because partial unfolding of proteins with low to moderate amounts of denaturant or heat produces aggregation, whereas complete unfolding results in reduced or no aggregation at high denaturant concentrations or temperatures.\textsuperscript{116} Natively unstructured proteins do not inherently aggregate rapidly and these too must be provoked to do so by partial ordering.\textsuperscript{117} It seems reasonable, based on the understanding gained from protein folding studies, that a critical density of hydrophobic contacts would be required to stabilize any condensed state regardless of whether contacts stabilize the native state or are involved in aggregation.\textsuperscript{63} In many proteins, it is apparent that the appropriate conditions for aggregation are met in partially unfolded states, but experimental requirements do not permit determination of whether aggregation proceeds solely from a partially unfolded state or whether the fully unfolded form also aggregates rapidly. In some cases, however, completely unfolded forms of proteins do not appear to aggregate. When high concentrations of denaturants are used, the unfolded state does not readily aggregate, but this appears to be due to competition between the chemical denaturant and protein molecules for interaction. Theoretically, if transition to the fully unfolded state could be accomplished instantaneously and enough thermal energy is imparted to completely disrupt the structure (which also seems to be required for aggregate formation), then the heat-denatured state should be less prone to
self-association as well. The folding and unfolding processes of globular proteins have been studied extensively using proteins that unfold reversibly and obey two-state kinetics.\(^{118,119}\) Aggregation has been considered with respect to the parameters derived from these models, but because aggregates are difficult to study directly and in most cases are irreversible or are only very slowly reversible, these models are unique and continue to be revised. Aggregation models are based on forward rates of association at varied concentrations and conditions to form large aggregates and intermediates that precede the final stable form. Because the morphology and size of fibrous aggregates are easier to correlate with specific changes in structure than large amorphous agglomerates, most models have been developed based on data from amyloidogenic peptides and proteins.

### 1.4.1 Monomer Conversion

Templated assembly (TA) was proposed soon after the discovery of prion proteins to explain their mechanism of infectivity. A dominant conformational change in one monomer (α) provokes rearrangement in native monomers (α’) by acting as a template (α\(_2\)) from which the aggregate grows.\(^{120}\) A favorable energy difference driving α’ to α is proposed, but the barrier to conversion would be large for the wild-type protein under normal physiological conditions. This makes formation of the dimer highly improbable. The existence of α\(_2\), however, would then lower the energy barrier for a subsequent conformational rearrangement of α’, allowing α\(_2\) to act as a template for propagation. The spontaneous onset of disease and conversion could be explained by any perturbation that lowers the initial activation barrier to permit the formation of α\(_2\), such as a mutation or a change in the environment. This model predicts that the rate of assembly will be directly proportional to the concentration of α\(_2\) and will not depend on the concentration of α’ over a wide range unless α\(_2\) is in excess. Based on the assumptions of this model, α’ will always be in great excess over α\(_2\) since the formation of the dimer is rate limiting. Because prions are infectious, they must remain soluble in aqueous solution and must cross lipid membranes. This duality suggests that the existence of a fluid equilibrium between multiple states is necessary for transmission. The existence of multiple equilibria in disease states should apply equally well to non-disease proteins in concept, although the structural transitions may not occur in such uniform ways to generate highly organized fibrils. The monomer-directed conversion (MDC) model is based on the same concept involving a conformational change in aggregate assembly.\(^{121}\) In this model, however, the aggregate does not act as a template. Instead, the barrier to structural rearrangement is minimal and assembly occurs as a consequence of frequent conversion to the aggregation-prone conformation. The rate of formation of aggregates increases with increasing concentration of protein in this model because the probability of productive collisions increases.
1.4.2 Oligomeric Intermediates

The nucleation–polymerization (NP) model was developed to account for the lag phase preceding the growth of insoluble fibers from untainted protein samples. The NP model resembles crystallization processes in which nucleation is the rate-determining step and is followed by a more rapid growth phase.\(^\text{122}\) The formation of nuclei depends primarily on the initial concentration of soluble protein, such that the rate of formation increases exponentially with increasing concentration.\(^\text{123}\) The rate also depends on the equilibrium of conversion between the native and aggregation-prone states. If the kinetic barrier is easily surmounted, nucleation will proceed more rapidly than if the barrier is more difficult to cross. The upshot of this model is that there is a critical concentration at which stable nuclei are able to form and persist long enough for growth to occur. The oligomeric nucleus must remain associated longer than the effective rate of diffusional encounter with monomers having an aggregation-prone conformation. The growth phase is predicted to be constant at monomer concentrations in large excess of nuclei and to taper off as the concentration diminishes through incorporation of protein into the aggregate. If preformed nuclei are introduced into a solution of monomers, they act to seed growth, eliminating the lag phase. This model was developed based on \textit{in vitro} studies involving the amyloid-forming peptide PrP_{96-111M}, which complies with all aspects of this kinetic model of aggregation. Although the results of these studies support the NP model and full-length PrP forms amyloid plaques \textit{in vivo}, it is evident from other studies that neither the TA nor the NP model explains fully the mechanism of PrP\textsuperscript{C} conversion to PrP\textsuperscript{Sc} with regard to both infectivity and plaque formation.

The nucleation conformational conversion (NCC) mechanism employs elements of both TA and NP to explain the transformation of amorphous aggregates into highly ordered aggregated structures. Experiments conducted using the NM (glutamine-rich N-terminal and highly charged middle regions) of the yeast prion protein Sup35 showed that the kinetic rates of assembly were inconsistent with both the TP and NP models.\(^\text{124}\) The rates did not vary much with protein concentration over a broad range and were limited to a maximum rate at very high ratios of monomer to nuclei. In NCC, like the NP model, the kinetic barrier and rate-limiting step in aggregation is oligomerization rather than conversion of the monomeric conformation. The oligomeric species and monomers exist in equilibrium, but the concentration of oligomer dictates the rate of fibril formation, as opposed to the monomer concentration directly determining the rate of further growth. The growth phase proceeds through the fusion of preformed nuclei, generating larger aggregates as a result of interaction between oligomers. This mechanism implies that the conformation of the monomer in the context of the oligomeric structure has a lower barrier to further aggregation than does the soluble monomer. In this model, the oligomer must be accessible from the normal conformational milieu of the native protein or a partially unfolded state in which a specific segment becomes
exposed, which initiates aggregation. Perturbations, such as mutation or changes in pH, which increase exposure of such nucleation sites, have been identified for several proteins. PrP aggregation appears to depend on residues 96–111, which is buried in the native structure, and many peptides derived from the core regions of various normal and disease-causing proteins aggregate rapidly in solution. The increased rate of aggregation of the peptides would therefore be a consequence of removing the protection afforded by the native fold.

1.4.3 Nucleation in Protein Folding

Because proteins usually fold cooperatively, there is a need to understand the interplay between local and distal interactions that synergize to develop a condensed state. One idea that has been presented is that a local structural nucleation site (or sites) exists around which the rest of the protein coalesces. A foldon is a short, contiguous amino acid sequence with a preference for a particular conformation. Foldons were proposed to act as nucleation sites within protein sequences that aid in the specificity of folding. To achieve cooperative folding following two-state kinetics, the foldons would have to overlap. If the foldons are separated in space, then multistate kinetics will be observed as cooperativity breaks down and intermediates accumulate while the separate foldons search for preferential modes of interaction that lead to the native conformation. The foldon concept fundamentally implies that sequence-encoded information dictates folding. Cooperativity would occur largely through predisposed backbone angles with concomitant compatible side-chain interactions both locally and distally. With respect to this idea, partial unfolding of a protein may lead to increased exposure of regions that retain such structural features, and these foldon-like segments may then nucleate intermolecular assembly.

The diffusion–collision (DC) model presented by Karplus and Weaver is based on the formation of microdomains within the protein structure, such as transiently stable helices. The kinetic process of folding is dictated by the diffusion and collision of microdomains, such that productive collisions result in the coalescence of higher structure. In most applications of this model, only native collision-coalescence events are permitted. Nonnative interactions were allowed, however, in a study by Beck et al., when modeling a four-helix bundle to characterize transient nonnative collisions, potentially indicative of misfolding events. DC and foldons share the common idea that preformed structural elements or tendencies are encoded in the sequence, but DC differs from the foldon concept in that folding need not be fully cooperative and is quantified using a molecular modeling approach. The existence of microdomains simply restricts the folding pathway options, which permits the formation of off-pathway associations. This view is consistent with observations from in vivo systems because it accommodates the possibility of off-pathway and slow-folding intermediates, which are formed often in proteins and require
assistance from chaperones to achieve their native conformation. No particular kinetic mechanism is inherent in the DC model of protein folding, and as such, it is amenable to characterizing proteins that have great diversity in the rate of folding and in the number of paths available regardless of whether intermediates form. The microdomains arise from fundamental features of the sequence, and the collision between domains depends on the distances separating the domains. With this approach, mutations can be introduced to assess easily the effect on folding rate and the formation of intermediates. A similar concept may play a role in the aggregation of proteins, in which aggregation “hot spots” are unveiled as a protein unfolds. Peptide sequences derived from proteins that aggregate are prone to rapid assembly when examined in the absence of the full, folded sequence.

1.4.4 Domain Swapping

Domain swapping among dimers has been observed in a number of proteins. In some cases, the structure of the fold is virtually identical to the native monomer forming a closed interface, while in others, the interaction results in a distinctly different conformation for the participating residues. Proteins that swap domains often aggregate further, and it has been proposed that domain swapping is an intermediate on the way to the formation of larger oligomers. The phenomenon has been investigated in detail for human cystatin C (HCC), where mutation of L68 to Gln causes amyloid angiopathy. This cysteine protease inhibitor may function as a monomer, but when destabilized by the L68Q mutation, it associates into dimers. The mutation occurs in the third beta strand (β3) and disrupts packing interactions by introducing a polar moiety into an otherwise hydrophobic space. Consequently, the energy barrier between the native and unfolded state is lowered, allowing the first two beta strands (β1 and β2) and alpha helix (α1) to cross over and interact with the sheet from the partner domain in the swapped configuration (Fig. 1.6). The original architecture is retained on each half of the symmetric dimer, but a new “open interface” is created by the bridging interaction between the domains. The dimer suffers from the same packing problem as the monomer but overall is stabilized thermodynamically compared to the native state by the additional β-strand H-bonds in the linker region (βL). It has been proposed based on an examination of other domain-swapped dimers that the open interface is responsible for further aggregation. It is plausible that in HCC, conversion to larger oligomers may emerge at the open interface of the domain-swapped state because of the same packing defect, which concomitantly diminishes interaction between residue 68 and α1 via Y34 and V31 from the partner molecule. Movement of this helix would be a prerequisite to the formation of an amyloid cross-β structure, which HCC forms. Because α1 and β2 are more flexible in the L68Q mutant, disulfide bonds were engineered into this protein to anchor α1 and/or β2 to the main sheet where β3 and residue 68 are located. Each set of cysteine mutations stabilized the monomer, pre-
Figure 1.6. Domain-swapped dimer structure of cystatin C. The fold of chicken cystatin (PDB entry 1CEW), defining the topology of this class of proteins (top). Domain-swapped dimer of HCC in a view similar to (middle) and in a perpendicular orientation (bottom) emphasizing the beta sheet in the domain switch region (open interface) and the site of the L68Q mutation (red dot). From Reference 64. See color insert.

vented domain swapping, and greatly diminished amyloid fibril formation, indicating the domain-swapped dimer is an intermediate that propagates aggregation. Domain-swapped HCC subsequently forms metastable oligomers having a doughnut-like morphology with uniform outer and inner diam-
eters of 13.4 and 2.7 nm, respectively. The oligomeric state (or states) could not be accurately defined because they are not amenable to analysis by gel filtration, but they appear to be 200–300 kDa by sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE). The oligomeric species seems to be a metastable intermediate involved in fibril formation because it converts into fibrils much more rapidly than its monomeric and dimeric precursors.

An alternative mechanism by which fibrils form directly from domain-swapped dimers without occupying globular intermediates requires that domain swapping produce an open-ended oligomer in which additional monomers can assemble at both ends of the structure. This so-called runaway domain swapping inherently bears the requirement of high sequence specificity and was confirmed to exist using T7 endonuclease I (T7EI). Cys residues were engineered into the T7EI sequence such that intermolecular disulfide bonds would specifically result when embedded in the proposed organization. When aggregation was allowed to proceed under reducing conditions, wild-type protein-like fibrillization occurred. Domain-swapped dimers pre-stabilized by oxidation did not proceed into fibrils. Addition of increasing amounts of reducing agent revealed an equilibrium between two dimeric forms, only one of which was able to form fibrils, suggesting that a conformational change in the domain-swapped dimer precedes fibril growth (Fig. 1.7).

![Figure 1.7](image_url)

**Figure 1.7.** Schematic model for the fibril formation of T7EI L19C/S95C, illustrating the need for a structural transition from the dimeric state to generate the fibrillar arrangement. Each subunit is colored either in light or dark gray. Black dots represent disulfide bonds, and SH represents free cysteine. In dimer I, both disulfide bonds are intact and the protein is locked in close-ended dimers and is unable to form fibrils. In dimer II, half of the domain-swapped molecule is unlatched by the reduction of one of the two disulfide bonds. Upon incubation at 37°C, dimer II changes to an open form (dimer II*), in which half of the dimer opens up, exposing the interface that remains protected in dimer I. Open-ended dimer II* readily fibrillizes via runaway domain swapping. The hinge–loop region of the domain-swapped protein forms a zipper spine in the fibrils. Reproduced with permission from Reference 132.
From the studies described in this section as well as numerous others, it is clear that aggregates can have well-defined structures and can participate in specific interactions involving hydrogen bonding between beta strands. What remains to be explained is the selective basis for and the mechanisms by which a polypeptide achieves the native fold or an aggregate-prone intermediate conformation. While coordinated unfolding at specific sequences and association is required to generate fibrillar aggregates, amorphous aggregation may arise from heterogeneity among the unfolded population in which associations between several states have approximately equal free energies. This could permit an irregular macroscopic organization to develop as a result of intermolecular interaction at multiple sites.

1.5 MODELS OF PROTEIN FOLDING

1.5.1 Classical Models Based on Chemical Equilibrium

The earliest models developed to explain protein folding were based on chemical reactions at equilibrium, as defined by kinetics. The simplest conception envisioned a direct conversion between the unfolded (U) and native (N) state: U ↔ N. Inherent in the chemical equilibrium view is the idea that the protein has a single path by which U and N interconvert. In this case, conversion involves a single step with a free energy barrier separating them. Folding must be completely cooperative to fit this model. Indeed, many early observations made concerning protein folding and unfolding obeyed such two-state kinetics. These early studies were almost all performed on small globular proteins that fold cooperatively, and the results show that this model describes folding well for this type of protein, at least to a first approximation. Because the timescale for folding is often on the order of milliseconds, only the unfolded and native states were observed in equilibrium experiments, validating the model. With this line of thought, it follows that aggregation must proceed from the unfolded state, since the native form is uniquely folded. As such, aggregates were believed to be composed of random coils held together through nonspecific hydrophobic interactions. This model is vastly oversimplified based on the knowledge available today, which indicates aggregation usually arises from partially structured species that form organized interfaces.

During early folding experiments, a few proteins were observed to fold in a manner that did not obey first-order kinetics, suggesting that folding involved multiple steps and the formation of intermediates. Such intermediates had to persist sufficiently long and to accumulate in enough quantity to be detected on the experimental timescale. Some proteins were hypothesized and were later shown to undergo isomerization at Pro residues or to be altered by the incorporation of a metal or heme group, slowing the approach toward the native state and allowing the observation of a specific intermediate.\footnote{118} As a result, the model was amended to include a limited number of
intermediates in series along a single pathway: $U \leftrightarrow U' \leftrightarrow (U'' \leftrightarrow U^n) \leftrightarrow N$. This multistate model permits the formation of on-pathway intermediates along the reaction coordinate toward achieving the native conformation. More extensive analyses have shown that these intermediates largely resemble the native form with minor deviations, implying that they are late-folding or early-unfolding intermediates. Because tools were not available to detect states that did not accumulate, the formation of transient intermediates could not be unambiguously established until the advent of stopped-flow and rapid-quench methods. Disulfide bonds added additional complexity to this scheme by providing evidence for off-pathway intermediates. This was evident even in Anfinsen’s early studies in which refolding performed in the absence of a reducing agent led to a heterogeneous population of covalently cross-linked inter- and intramolecular species. These species remained soluble and were easily eliminated by refolding from the denaturant before allowing oxidation to occur. The mass action view of folding has limited options with which to describe off-pathway intermediates because it is based on a series of only a few sequential steps. Aggregates would then result from structures that are part of the native folding pathway. This model also does not lend itself to providing satisfactory explanations for the phenomena of nucleation and growth of aggregates.

![Scheme 1.1.](image)

Applications of classical models to protein folding often result in a poor fit to experimental data. Hydrogen–deuterium exchange (HX) experiments have revealed the coexistence of subpopulations of folding intermediates, indicating that multiple, diverse pathways exist for arriving at the final native structure from the unfolded state or vice versa. It was in part because of HX data that a new model for protein folding and unfolding emerged based in statistical mechanics. This is discussed further in the next section.

The ideal gas law has also been used as a limiting law tool for describing protein folding. This model is of value because it incorporates the concept that a continuous range of states exists during protein folding, opening up many options for how a protein transitions from the unfolded to the folded state. In this model, the boundary is set by the volume the protein is permitted to sample, and entropy is a function of volume ($V$) and depends on the number of residues in the chain ($N$) according to the following equation: $S(V) = N k_B \ln(V)$. The entropy of the unfolded chain increases with larger numbers of residues and is presumed to be $\sim 2.3 k_B N$. The residues are permitted to diffuse within the bounded space according to a potential function that drives
them away from the edge of the spherical boundary and toward the center of the sphere, promoting the condensed, folded state. Although this simple model was useful for beginning to explore the effect of entropy on the unrestricted and bounded states, the function used is often a poor mimic of reality, because different proteins of the same length do not necessarily condense to the same architecture, indicating that other factors exert major influence on how a protein folds. This does not invalidate the thermodynamic model but simply implies that details of the protein sequence are critical to understanding the mechanism and kinetics of folding as well. The poor fit of data to any model stimulates testing of additional parameters. The problem with using the ideal gas model is that there is no phase transition. Condensation requires attraction between atoms or molecules, as exist with nonideal gas systems. As confirmed by Anfinsen’s and others’ studies, hydrophobic interaction provides an attractive force to drive condensation during protein folding. The simple gas law-based approach to examining entropic effects was an essential first step in understanding protein folding and was the basis for the development of more sophisticated models that introduced attractive forces, such as hydrophobic interaction, and other restraints into the calculations. Another outcome of applying the ideal gas model to protein folding was the hypothesis that similar to phase transitions between the gaseous and condensed states, the formation of a nucleus may be required to seed the transition. The nucleation hypothesis applied to protein folding provides a possible mechanism by which Levinthal’s paradox is resolved. 137

1.5.2 Statistical Mechanical Models

A particularly important innovation in understanding protein folding was the idea that proteins do not follow a single path to the native state but rather have manifold options by which to arrive at the native conformation. Along with this idea came the consideration that proteins are unlikely to populate a single static state but instead exist as an ensemble of conformers. Such models are based on statistical mechanics, in which the minimum energy state corresponds to a collection of native structures and folding proceeds along any trajectory that is not energetically forbidden. 136 The result of this type of analysis produces an enormous number of different viable conformations that could be adopted along the way from unfolded states to the native structures. Moreover, a Boltzmann-like ensemble is obtained for the folded protein; it is made up of similar but not identical structures that define the native state. A major difficulty in the interpretation of such data emanates from the inability to observe individual conformers experimentally. The most detailed view that can be obtained currently is with solution NMR, which is useful in probing the conformational dynamics of a protein and for observing even small sub-populations (>1%) within the sample. 139 In practice, it can be quite expensive and time-consuming to acquire and interpret this type of NMR data. As such, computational analysis provides an alternative, efficient way to explore the
mechanisms behind protein folding and interactions, including aggregation. In silico experiments are easily modified to assess the effect of mutations and already are providing very useful insights concerning the mechanisms of folding and misfolding. The value is maximized when the results of simulations are used to inform experiments and vice versa.

The computational approach involves calculations reliant upon several physicochemical parameters, corresponding to a highly multidimensional space of the function $E(\Theta_1, \Theta_2, \Theta_3 \ldots, \Theta_n)$, where $E$ is the energy derived from an equation involving one or more parameters ($\Theta$). The output is often simplified for visualization in the form of an energy landscape. The axes of the two-dimensional landscape are entropy on the $x$-axis and internal free energy on the $y$-axis. The profile for a protein that undergoes fast folding looks much like a funnel with rough edges (Fig. 1.8). The better the protein complies with the classical two-state cooperative folding model, the smoother the funnel appears. The landscape is often rendered as a three-dimensional plot in which the $z$-axis is the internal free energy of a chain conformation and the axes in the transverse plane reflect specific parameters such as torsion angle energies, hydrogen bonding, and van der Waals contact, defined by $E(\Theta_1, \Theta_2)$. Each point on the energy surface corresponds to a conformation of the protein. In general, the appearance of the landscape reflects kinetic differences between the conformational states, such that points at similar latitudes correspond to states that are equally probable. The slope of the surface is indicative of the rate of folding, in which steep edges indicate fast rates. Importantly, the results of these models can be evaluated by comparison to measured energy values obtained from various experiments. There are many approaches used to calculate the energy, and the variations are derived from observations made

Figure 1.8. Two-dimensional energy landscape diagram of entropy versus energy. A semi-smooth, single funnel represents a typical protein folding landscape (left), whereas the landscape depicting aggregation is more complex and has multiple, well-separated minima corresponding to distinct stable structures (right). From Reference 141.
using a wide range of experimental tools that help further refine our understanding of the parameters that influence folding and aggregation.

While protein folding landscapes typically are funnel shaped, theoretical studies of protein aggregation produce much more complex landscapes with many traps\textsuperscript{141}. The degree of complexity reflects the kinetic and equilibrium behavior of proteins at high concentration, in which native contacts compete with alternative intermolecular interactions.

1.5.3 Computational Models

Technological advances that enhance computer processing speed and data storage capacity have enabled the ability to model protein folding and other interactions. Structure-based approaches have provided the ability to test the influence of individual parameters on protein folding, such as the strength of hydrophobic and electrostatic interactions. Results can be validated and the accuracy of the result assessed by comparison to experiments. The validation of the approach and the role specific parameters exert on folding has made ab initio calculations possible to perform to assess the folding pathways of proteins and to identify intermediate states based on first principles without the need for a priori knowledge of the structure. Computations are based on parameters observed through experiment to affect protein folding rates guided using an energy function and restraints derived from experimental values. Ab initio calculations are not practical for use on systems with a large number of atoms, and for this reason, models have been simplified in a number of ways to make computation feasible.

One of the earliest models used to investigate protein folding using a computational approach was the HP model in which residues were designated as either hydrophobic (H) or polar (P).\textsuperscript{143} The protein is placed into a two-dimensional lattice such that each residue occupies a position on the lattice. Residues are arranged randomly in an unfolded state with their connectivity preserved. Folding is dictated solely by interaction between H residues (HH) due to the assignment of a favorable energy term, while HP and PP interactions are ignored because no energy term is assigned. Energy is introduced into the system, prompting the residues to rearrange through movement to open lattice positions. An exhaustive search is conducted to identify all possible conformations and their associated energies. Folding using this approach shows multiphase kinetics in which collapse occurs rapidly followed by slower steps corresponding to local rearrangements. Rearrangements are slower because this requires energy barriers to be surmounted. Based on comparisons to experiments, this model reproduces well general features of protein folding, likely because it provides a complete characterization of the energy landscape and reveals interactions that are preferred during the folding process. The landscape is rough due to the presence of numerous kinetic traps caused by use of a single parameter. The HP+ model incorporates an energy penalty for nonnative interactions, speeding the process of folding and minimizing the
AGGREGATION OF THERAPEUTIC PROTEINS

depth and number of kinetic traps, making the landscape smoother. This method cannot be applied to a protein with unknown structure. Proteins display non-Arrhenius dependencies on temperature during folding and unfolding experiments. By incorporating the temperature dependence of hydrophobic interactions derived from oil/water partitioning experiments, these simple models were able to reproduce qualitatively non-Arrhenius behavior and to attribute it to the temperature dependence of the hydrophobic interaction. This model was also able to explain chevron plots and observed deviations from ideal behavior (linear dependence of rate on denaturant concentration). The results underscore the central role of hydrophobicity in forming the condensed state.

The lattice method has been modified for use in three dimensions and with expanded alphabets to attain higher specificity. Additionally, more sophisticated terms are incorporated to account for polar interactions. When full alphabets with all 20 amino acids are involved, the Miyazawa–Jernigan (MJ) amino acid interaction matrix is typically used to define the interaction potentials between residues. The MJ potential is derived from empirical data, which presumably makes it more accurate than simpler potential functions like HP when used in lattice model calculations. Off-lattice models have also been developed, and this approach is more realistic and more accurately mimics protein folding. Specifically, it is better suited to predict the structures and assembly process of aggregates. It was used to analyze a 46-residue fragment of staphylococcal protein A that forms a three-helix bundle. The inclusion of all side-chain atoms in the model produced dynamics closely resembling experimental results, and this positive outcome was attributed to realistic side-chain packing. Constant temperature, simulated annealing (SA), or replica exchange protocols may be implemented using molecular dynamics (MD), discrete molecular dynamics (DMD), or Monte Carlo simulations, depending on the time frame of interest in the study and the parameters used. DMD is a good choice for aggregation analyses because it is able to explore significantly longer time frames, which is necessary for examining the formation and rearrangement of intermolecular interactions in assembly processes.

Go models were developed for use in the early stages of computational modeling of protein folding and are still useful. They rely on pairwise summation of contact energies derived from the known native fold. The method has been used extensively to characterize and test how individual parameters affect folding of proteins to the native conformation. This model cannot be applied to proteins or aggregates where the high-resolution structure has not been determined. Distance cutoffs are applied to define contacts and a binary (−1 or 0) approach is used to calculate the energy. It can be used with any bead model to further reduce the cost of the computation. The primary advantage of Go is that there are large energy gaps between native and other states such that folding proceeds rapidly in a cooperative manner. Because Go relies on known native contacts, aggregates that are based on...
nonnative contacts cannot be observed using Go models.\textsuperscript{149} They are, however, amenable to producing aggregates in which native contacts are preserved. An expanded version of this model was used to identify early stable local elementary structures (LESs) involved in the folding process. Pairwise potentials were added based on $\Delta \Delta G$ measurements from mutants that affect the native conformation.\textsuperscript{150} The LESs were found to provide the strongest interactions at the early stages of folding. These are needed to surmount the transition state barrier and to limit sufficiently the number of possible conformations to accomplish rapid, cooperative folding. This finding is consistent with the microdomain and foldon concepts of modular assembly and nucleation theories of folding.\textsuperscript{127,151}

\subsection*{1.5.4 Application of Coarse-Grained Models}

Protein folding and aggregation are analyzed and predicted based on knowledge of parameters that affect these phenomena \textit{in vitro}. Terms that describe intramolecular forces affecting the protein and solvent effects on the protein are included. Specifically, energy terms related to hydrophobicity, backbone ($\phi, \psi$) and side-chain torsion angles ($\chi_1, \chi_2$), steric strain, electrostatic attraction and repulsion, dielectric constants, hydrogen bond angles and lengths, aromatic interactions, and van der Waals contact and overlap may be included. Each is applied based on an energy function, often a continuous (Lennard–Jones) or a square well potential. The more constraints used in performing the calculation, the more time-consuming the computations become. Likewise, the computational time increases as more atoms are involved, such that all atom simulations are very computationally expensive and intractable for systems with greater than a few tens of atoms. Consequently, simplified \textit{coarse-grain} models have been used extensively to examine proteins because they greatly increase the speed with which computations are performed.\textsuperscript{152,153} Despite being greatly simplified, the results typically agree with experimental data, indicating the value of the simplified approach. Also, for this reason, solvent is not explicitly defined in many calculations, and instead, an implicit value for solvent is applied uniformly to the nonprotein space that reflects the average properties of water. It is more feasible to include explicit water in computations involving short peptides, where the number of atoms is relatively small. This offers the advantage of providing information about interactions between water molecules in the hydration sphere and the peptide that may influence the peptide’s conformation. In the past few years, coarse-grained models have been used increasingly for the examination of protein–protein interactions and aggregation of peptides because they reduce the cost of calculations sufficiently to permit the study of molecular systems with very large numbers of atoms. These systems are difficult to study experimentally and the information provided by computational methods has enabled the design of tractable experiments to test and refine hypotheses about the structure and mechanisms of protein aggregation phenomena.
Coarse-grained models are simplified because they replace individual atoms with a single bead that reflects the composite properties of the group it represents. The group can be defined in a number of ways, and typically, bead models employ one to six beads per amino acid instead of using all atoms. The coarser the model, the more important the accuracy of the parameter-based force field is in interpreting the result, and the more general the conclusions must be about the system. One-bead models reflect the average properties of an entire amino acid and place the bead at the centroid of the residue or $\text{C}_\alpha$ position. A small number of parameters are included in one-bead calculations that reflect the general size, geometry, and properties of the amino acids. The force field parameters are simplified to correspond with the one-bead approximation. This model is useful for comparing differences in the mechanism of folding between similar proteins where the topology is known, but not for de novo folding. The simplicity of the approach requires that the angle and dihedral terms be known, which means a high-resolution structure must be available. The simulation is then guided by terms reflecting the side-chain contributions to folding. The method has provided increased understanding about how specific side-chain interactions influence protein folding. These models lack the molecular detail needed to assess the contribution of specific backbone interactions and, as such, do not provide sufficient detail to explain the mechanism or structural differences underlying the formation of aggregates.

Two-bead models have the advantage of providing information about backbone geometry and hydrogen bonding. One bead corresponds to the backbone atoms and a second to the centroid of the side chain. Separating the two groups increases the accuracy of the prediction by making it reasonable to incorporate additional terms into the force field, including angle–dihedral angle correlations or secondary structure propensity. The addition of a third bead to longer side chains further improves accuracy by allowing more flexibility and distinct features to behave independently in the simulation. The additional beads greatly improve folding studies, but the model still has limited utility for examining aggregation because hydrogen bonding and backbone angles are crucial determinants, and these parameters remain too generalized in this approach.

Models with a minimum of four beads are needed to define explicitly the backbone atoms and to examine hydrogen bonding. Three of the beads are assigned to backbone atoms (N, $\text{C}_\alpha$, and CO), and side chains are represented by one (commonly $\text{C}_\beta$) or more beads. The array of four- to six-bead modeling approaches that can be performed is expansive, and it is beyond the scope of this chapter to cover them in detail so this description is limited to examples pertaining to aggregation. A four-bead model was used to examine dimerization of Aβ(1-40) and (1-42) peptides. Nine structures of planar beta-sheet dimers having equal free energies were observed, and no significant difference in energy was found between the two peptide sequences. The result demonstrates the ability of this model to generate the expected secondary struc-
ture characteristics of fibrous aggregates but, as the authors point out, also obviates the need for improved parameters to help differentiate between these structural isoforms since all atom simulations are intractable. Polyalanine peptides have been modeled under different conditions, and the results demonstrate the critical role of solvent, temperature, and peptide concentration on the distribution of monomer and aggregate structures obtained during the modeling process.\textsuperscript{157,158} Refinement of the parameters appears to be a significant challenge, which is just beginning to be addressed.

Only a few computational studies have been published concerning protein aggregation. The aggregates studied involve small peptides known to form fibrillar structures. The first employed mid- to high-resolution modeling methods and obtained flat beta-sheet structures.\textsuperscript{147} A lower resolution approach using one degree of freedom for the backbone dihedral and two intramolecular dipoles also was shown to generate fibrils.\textsuperscript{159} The major deficit of these approaches is that neither was able to reproduce the twist observed in fibrillar sheets. A mid-resolution approach using a simplified alphabet was developed to rapidly model a peptide aggregation that reproduces the twist associated with fibrils as well as the known interstrand and intersheet associations.\textsuperscript{96} This approach retains most of the backbone degrees of freedom, which are necessary for hydrogen bonding and chirality, and includes a single bead for the side chain to account for intersheet associations. The peptide contained alternating HP known to encourage beta structure and included oppositely charged residues at the termini. This model revealed the formation of a single hydrogen-bonded sheet (tape) early in the simulation, in which the strands formed an antiparallel arrangement. Double tapes emerged from the single tape only after a “critical nucleus” was obtained. In this study, the strength of the hydrophobic interaction and the polarity of the charged side chains were modulated. Double tapes formed solely from association of hydrophobic surfaces, which is consistent with the experimentally observed cross-\ensuremath{\beta} structure of fibrils. Higher-order tape formation requires the association of polar surfaces, and these species were only observed to form when the polarity of the charged residues was reduced in the simulation. The kinetics of fibril formation in this model are consistent with a nucleation–growth mechanism, in which a single tape of critical length serves as the nucleus and electrostatic effects influence nucleation. This model clearly provides a good platform for further investigation of factors that contribute to nucleation, growth and the structure of aggregates.

1.5.5 Information Theory (IT)

IT is a network analysis model focused on quantification of information and relationships based on the probability of an event’s occurrence. Originally, IT was developed in 1924 at Bell Labs by Harry Nyquist for the compression and reliable transmission of data by electronic communication. This fundamental approach relies on simple concepts and has been expanded since then for use
in a wide variety of applications, including molecular networks. The goal is robustness or to maximize efficiency of information transfer without compromising the quality of the output on the receiving end. With electronic data, this means determination of the minimum number of bits necessary to achieve an accurate and correct interpretation of the information during transmission over noisy, lossy lines. Line noise introduces uncertainty into the transmission and is considered to be equivalent to the entropy of the system. The efficiency of information transfer emerges from the use of probability functions in this model, because communicating the mean and distribution that describes a relationship more efficiently summarizes the system compared to transmitting the entire set of individual measured values that compose the system separately. Both approaches generate a similar interpretation following transmission, but one is more compact than the other. The two main features of quantification are entropy and mutual information. Entropy reflects the randomness of a variable, and mutual information describes the relationship between variables. In thermodynamics, the Boltzmann distribution function and Gibbs free energy equation encompass the base mathematical components of IT.

IT-based approaches have been applied to predict protein folding and aggregation. They parallel the simple thermodynamic model that based entropy on chain length such that molecular details of the proteins are not included in the model. Instead, general principles and macroscopic observables that contribute to association are used for the calculations. Early on, an approach that accurately describes the primitive hydrophobic effect of solutes in water was developed by Pratt and coworkers. The model predicts the probability of observing a solute-sized cavity in water based on the density of water and on the radial distribution of the central oxygen atoms in the solvent molecules. The two-moment model emphasizes the thermodynamics of hydration of the more hydrophobic solute and the association between hydrophobic entities, simply modeled as hard spheres. The model quantifies the hydrophobic effect using the excess chemical potential ($\Delta \mu_{ex}$) of the solute as the variable, because it reflects the probability ($p_o$) of finding a cavity of a given size/shape in the solvent according to the function $\Delta \mu_{ex} = -k B T \ln(p_o)$. This simple approach reproduces accurate plots of cavity-formation thermodynamics and the hydrophobic effect when applied to two simple, small-molecule model systems. A study of methane in water was performed, yielding the same profile as the more complex Smith and Haymet model. Application of the approach to $n$-butane generated lower free energies for the more compact cis and gauche conformations compared with trans, paralleling the results obtained in a more complex analysis by Begelov and Roux.

In one application to protein aggregation, Kumar et al. used a similar approach to characterize the interactions between neurofilaments (NFs) in axons. NFs are composed of monomers that form a central core along the long axis and a brush-like structure on the outside that is composed of a largely disordered sequence having an extremely high percentage of charged
residues. Association of NF was modeled using IT based on experimentally derived values. The cross-sectional radial-spatial distribution function (RDF) of NF in axons describes the density of NFs around a single central NF, providing the probability distribution \( (g(r)) \). The potential mean force \( (u_{MF}) \) can then be determined for the average interaction potential over all configurations and orientations of NFs in the system, according to \( u_{MF} = -kT \ln g(r) \). A central NF was randomly selected and the number of particles counted within a defined spaced was used to determine the occupancy probability distribution (OPD). Fitting to the IT model was then accomplished using the RDF and OPD data. Other modeling approaches require that interparticle distances be known or estimated. Because the measurement is not possible in the NF system, the IT model is advantageous because this data is not needed for the calculation.

Several IT-based models have been used to assess aggregation propensity for proteins using the amino acid sequence. The Thornton group used an approach based on the propensity of each individual amino acid to form a β structure to predict the preference of a peptide to then form pairings between strands that would lead to parallel or antiparallel beta-sheet structures.\(^{162}\) The method was able to identify the correct register less than 50% of the time but was able to distinguish aggregation-prone from random sequences. Since it has been established that hydrophobicity is directly related to the amount of solvent-accessible surface area for polar and nonpolar groups alike, another approach to understanding aggregation was based on prediction of the solvent accessibility of amino acids in the context of different sequences.\(^ {163}\) Calculations were performed using values for each individual amino acid alone and for each in the context of a 17-residue peptide. The results confirmed that the side-chain solubility is strongly influenced by local context, and a revised hydropathy scale was presented to account for neighbor effects on solubility. Like the two aforementioned approaches, most studies have relied on small, hand-selected data sets in order to focus on analyzing a specific type of interaction. A systematic, high-throughput method was developed by Ofran and Rost to sort the very large structural data set available in the PDB to identify differences in composition correlated with distinct types of protein interfaces.\(^ {164}\) The approach confirmed findings from the previous studies on smaller data sets and also revealed the presence of six types of interfaces. The distinctions reflect the probability of interaction, which is derived from amino acid contact preference and interatomic distance, as determined by proximity within the polypeptide sequence and through covalent (obligate) attachment between chains. Homo- and heterochain interactions were found to have different bases for association, as was also noted among transient and obligate partners. The composition of amino acids within each type of interface was unique among the six classifications. In all types of interfaces, however, lysine, serine, alanine, and glycine were underrepresented, while arginine was overrepresented. Hydrophobic residues as a class were found with approximately equal
frequency in all types of interfaces, but the specific apolar residues present varied between interface types. For example, on average, leucine and valine are abundant in protein interfaces, but the percentage of each varies dramatically between unique types of interfaces. Hot spots for association were previously reported to have an abundance of tryptophan, tyrosine, and arginine and to lack serine, threonine, valine, and isoleucine. Separating the interface into two regions, the core and rim, leads to further distinction in the involvement of individual amino acids. While these general observations were affirmed in this study, more specificity among residues was gleaned from each type of interface. For example, on average, tryptophan is overrepresented in protein interfaces, but it is extremely underrepresented in homocomplexes.

Mintseris and Weng further simplify the IT modeling approach by employing a five-letter alphabet. They grouped the residues by type to more rapidly examine the influence of specific properties of the side chain on folding and interfacial association. Grouping reduces the amount of mutual information, making the calculation more efficient. Analogous to coarse-grained bead models, the groupings encompass contributions from the main chain and from hydrophobic, positively charged, negatively charged, and polar side chains. This study predicts that hydrophobicity within a sequence has a greater influence on monomer folding than on aggregation and also that polar and electrostatic interactions are important factors in interface formation.

The major advantage of an IT-based approach is the simplicity of the model, which remains capable of producing results analogous to more complex approaches. Alternative or additional physical insights can be incorporated directly into the model, allowing it to be applied to the investigation of more sophisticated systems like proteins. Some additional applications of its use for the study of proteins not discussed here include interactions of ligands at binding sites, the effects of mutations on protein stability, and examination of conformational equilibria of protein side chains.

1.6 INFLUENCES OF CHEMICAL ALTERATION ON AGGREGATION

Proteins are subject to both enzymatic and nonenzymatic changes to their chemical composition, which alter their conformation and/or stability. In vivo, posttranslational modifications frequently control activity and protein–protein interactions. This can be accomplished by oxidation, reduction, proteolysis of the polypeptide chain, or other chemical modifications that alter the surface properties and/or conformation of the protein. Spontaneous changes, such as deamidation, were proposed to function as internal timers signaling that the protein should be actively repaired or degraded. Similarly, spontaneous chemical changes to proteins in vitro can alter the conformational dynamics and can affect protein stability and aggregation. The implications of chemical
changes in proteins appear to be complex and depend not only on the amino acid affected but also on its sequential and structural neighbors as well. The major types of spontaneous chemical alterations that affect folding and aggregation are described in this section.

1.6.1 Length of the Polypeptide Chain

Cleavage of the protein backbone has been shown to alter the aggregation properties of both disease-related and therapeutic proteins. The Aβ(1-42) peptide is more prone to fibril formation than its slightly shorter counterpart Aβ(1-40). In this case, the presence of residues 41 and 42 increases the overall hydrophobicity of the peptide, and this is thought to be responsible for enhanced aggregation. Proteolytic cleavage of blood factor proteins promotes aggregation in clot formation. Here, it is the removal of a propeptide that initiates assembly. The clotting factor is protected from rapid aggregation because while covalently attached through the peptide bond, the propeptide region effectively interacts with the aggregation-prone region. Once the pro-moiety is cleaved, the peptide dissociates and no longer impedes intermolecular aggregation. Spontaneous hydrolysis can also occur during storage, often at Asp residues, and depending on the protein, this may promote aggregation.

Addition of residues to the terminus of a normal protein also may facilitate aggregation. The most prevalent examples of this are found in the polyglutamine tract extensions associated with diseases that form nuclear plaques. Modeling and in vitro studies suggest that an extension of a sufficiently long length increases the dynamics of the native protein such that partial unfolding results. In contrast, polyglutamine sequences alone assemble into fibrils and amyloids. A viable alternative proposal is that unfolding of the native protein ensues following aggregation of the polyglutamine segment.

1.6.2 Methionine Oxidation

Oxidation has been predicted to promote aggregation of proteins involved in amyloid plaques because oxidative stress is strongly correlated with protein misfolding diseases. Despite the fact that oxidation of numerous proteins embedded in amyloids is increased significantly, it appears that Met oxidation discourages amyloidosis. High-resolution NMR studies of Aβ peptides show that oxidation of Met35 to methionine sulfoxide substantially decreases aggregation rates. Spectral changes suggest a conformational adjustment occurs in response to oxidation that disfavors formation of the turn involving Gly37 and Gly38, thereby inhibiting self-association. Oxidation at methionine was also shown to disrupt fibril formation in the prion protein PrPSc. Interestingly, modification did not prevent aggregation entirely, since prefibrillar aggregates accumulated. In vitro studies on recombinant factor VIIa (rFVIIa) also show that oxidation of its methionine
residues fails to promote aggregation. In fact, conditions that stabilized rFVIIa against aggregation had increased methionine sulfoxide content. Oxidation of methionine increases the polarity and hence solubility of the hydrophobic side chain. In keeping with this, studies of these and other amyloid proteins indicate that the site at which oxidation occurs does not encourage aggregation through direct participation in the intermolecular interface but may do so indirectly by promoting unique structural transitions that are responsible for generating low-molecular-weight oligomeric aggregates and large fibrillar forms.

1.6.3 Covalent Cross-Linking of Strands

Oxidative modification of proteins can lead to covalent cross-linking. This predominantly occurs through the formation of disulfide bonds between Cys residues, but covalent attachment through lysine and aromatic side chains has also been observed.

The involvement of disulfide bonds in protein stability and aggregation appears to depend in part on the biological environment in which the protein is normally found. Evolutionary selection occurs within a specific environment and protein sequences have been selected to avoid aggregation in vivo. Extracellular proteins, which function in more oxidizing environments, often contain disulfide bonds that stabilize their native conformation. These proteins are often used as therapeutics and include insulin, growth hormones, interferons, and monoclonal antibodies. Cleavage of disulfide bonds by reducing agents leads to increased conformational mobility and uncondensed states that resemble MGs or random coils. Mixing of disulfide bonds generates nonnative structures, including covalently linked intermolecular species that facilitate aggregation. For example, in the absence of its metal ligand, an intramolecular disulfide bond forms in Cu, Zn superoxide dismutase (SOD1) that involves the Cys residues that normally participate in metal coordination. This leads to increased flexibility of a loop at the dimer interface and promotes aggregation.

Cytosolic proteins are buffered from oxidation in vivo due to the presence of excess reducing capacity, largely in the form of glutathione (GSH). As such, the vast majority of Cys residues in these proteins are in the reduced state. Oxidation rates depend on the pKa of the thiol, such that more easily ionized Cys are more susceptible to reaction. Solvated Cys residues typically have pKa values about 8.7, but the pKa can be perturbed by several orders of magnitude as a result of the local chemical environment developed in a uniquely folded state. Oxidation regulates the activity of some intracellular proteins. For example, the protein tyrosine phosphatase’s (PTPase) catalytic action relies on the existence of a reactive thiolate anion in the active site. Oxidation inactivates these enzymes and alters their conformation, facilitating their removal, possibly to prevent aggregation. Oxidative damage is strongly correlated with protein misfolding diseases, suggesting that the formation of
disulfide bonds may play a role in intracellular aggregation. Studies of the oxidation of proteins that do not contain native disulfides often report that disulfide formation leads to aggregation or changes in solubility. This parallels the effect of disulfide mixing observed for proteins that contain native disulfide linkages.

Covalent dimerization of α-synuclein following oxidation promotes aggregation. This protein contains no Cys and so attachment is achieved not through disulfide bonding but instead as a result of tyrosine oxidation to form dityrosine. Cross-linking of these aromatics promotes fibril formation. Selective Tyr cross-linking is achieved by employing the cytochrome c/H₂O₂ system, which has the added benefit of avoiding oxidation of other amino acids. Oxidation often is carried out using Fenton chemistry, which results in the modification of several residues, including Met, Cys, and His, but does not produce dityrosine. Depending on the rate at which oxidation proceeded and the type of oxidant used, aggregate morphology for α-synuclein differed. Such results indicate the importance of spatial and temporal parameters on the aggregation mechanism.

Covalent linkages form between strands in collagen matrices that involve modified lysine residues. The normal structure of collagen consists of a triple helix, which constitutes the base unit in a higher-order network that composes basement membranes in vivo. This organized aggregate forms a semipermeable barrier with a small pore size, which limits infiltration of macromolecules into the tissue. The tight packing arrangement depends on covalent cross-linking and association of two hydrophobic regions. A staggered antiparallel orientation develops between triple helices, organizing the structure into tetrameric fibers. Tetramer formation is driven by interaction between two stretches of apolar residues and is stabilized by regular, covalent attachment through two disulfide bonds and two lysine-derived aldimine bonds per strand. In the native matrix, aldimine linkages are found between specific lysine residues in the N-terminal region where hydroxylysine is found followed by a Gly-His/Glu-Arg sequence. The majority of lysine residues in collagen are modified to hydroxylysine or O-glycosylated hydroxylysine. These modifications sterically hinder association between helices on the outside face of the tetramer, which imparts both elasticity and strength to the polymer, supporting the function of the basement membrane. As the matrix ages, the number of glycosylated lysines increases and the extent of cross-linking increases. Glucose reacts with the ε-amino lysyl moiety and, through a series of steps, ultimately leads to the formation of a stable aldimine linkage. Exposure to increased concentrations of oxygen and glucose accelerates this reaction at other sites, causing structural rearrangements that increase the pore size and alter the permeability of the membrane. Although the details of the resulting structure are not fully described, available data suggest that not only more cross-linking occurs but that the positions of attachment also differ compared to the natural lysyl-aldehyde-derived linkages.
1.6.4 Deamidation

Deamidation commonly abrogates function and may promote partial unfolding, which enables aggregate formation. Degradation of asparagine is a spontaneous process, which shortens the shelf life of numerous protein therapeutics. The reaction typically leads to the formation of isoaspartic acid (IsoAsp) and Asp in a 3:1 ratio. IsoAsp is a beta amino acid, and the inclusion of an additional carbon in the backbone would prevent the direct hydrogen bonding required to support beta-sheet formation at the site of the modification. This effect would thus need to propagate aggregation through conformational changes or by negatively affecting surface properties, such as charge complementarity or hydrogen bonding. Studies of the amylin 20–29 peptide show that the native sequence does not aggregate, but amyloid formation results from the introduction of a small amount of deamidated species. The mechanism resembles the template-driven aggregation postulated by Prusiner for prion proteins, which proceed in a pH-dependent manner like that observed with deamidation. Aggregates formed more rapidly at pH 5.7 than 2.25, suggesting that the ionization state of the new carboxyl moiety plays a role in assembly. In other cases, the introduction of a charged moiety, as would be the case at pH 5.7 for this peptide, has been shown to interfere with aggregation. In this case, however, the overall charge on the peptide becomes neutral following deamidation, lowering the barrier to intermolecular association. Although currently no structural data are available, it can also be speculated that the position of the carboxyl moiety near the positively charge N-terminus leads to a conformational change induced by electrostatic interaction.

Glutamine also deamidates spontaneously, albeit 100-fold more slowly than Asn. This modification has also been shown to alter the aggregation properties of proteins. In the case of the two-domain γD-crystallin proteins, deamidation of Gln at the interface between the domains diminished thermal stability due to the emergence of a charged side chain in a hydrophobic interface. The charged moiety disrupts packing and increases exposure of the remaining apolar residues in the core. The larger accessible hydrophobic surface area then makes the protein more amenable to aggregation. In contrast, aggregation of α-zein is diminished by Gln deamidation. Native α-zein is highly insoluble due to a high hydrophobic amino acid content and subsequent extensive aggregation. Gln can be converted enzymatically by protein-glutaminase (PG) to Glu for *in vitro* analysis. Deamidation by PG enhances solubility of the α-zein peptide by drastically reducing its propensity to self-associate. The mechanism of action appears to be due to the development of negative charge, which is uncompensated, disfavoring intermolecular β interaction. This finding parallels a common approach employed by nature to protect edge strands in beta sheet-containing proteins from aggregation.
1.6.5 Proline Isomerization

Protein folding studies have shown that a common rate-limiting step in arriving at the native structure is isomerization of proline.\textsuperscript{118} As this is a spontaneous reaction, \textit{in vivo} native-like states receive assistance in finding their correct conformation from peptyl-prolyl isomerases. \(\beta_2m\) in its native monomeric state contains a \textit{cis}-Pro at position 32. The native form is highly stable at physiological pH, resisting aggregation even at high concentrations, but oligomers form readily in the presence of \(Cu^{2+}\). Cu binding induces a structural transition in which Pro32 converts to the \textit{trans} isomer, initiating aggregation.\textsuperscript{193} Mutations at this residue invoke inherently the \textit{trans} conformation of the backbone and facilitate aggregation. \(\beta_2m\) oligomers formed from the \textit{trans}-Pro and mutant proteins eventually convert to amyloid fibrils, as occurs in the dialysis-induced disease state.

Despite the previous example, understanding the role of Pro in protein aggregation is not straightforward. A study of naturally occurring beta sheet-containing proteins found that one common mechanism by which native proteins prevent aggregation at the edge of their beta sheets is the incorporation of proline at a strategic position terminating the exposed strand. The sheet is terminated because the Pro creates the maximum allowable twist to support appropriate H-bond interactions on the concave side of the strand while preventing such interactions on the convex side.\textsuperscript{62} Proline scanning mutagenesis has been used to identify regions within amyloidogenic proteins that promote aggregation. Substitution of Pro into A\(\beta\) revealed that residues 15–21 and 31–37 are critical determinants of aggregation as insertion within these regions impaired plaque formation. This approach has been used successfully in a therapeutic as well. Amylin, which is used to treat hyperglycemia in diabetics, forms amyloid fibrils in the pancreas. Variants of amylin in which Pro is substituted in positions 25, 28, or 29 lack the ability to polymerize and therefore do not cause amyloidosis.\textsuperscript{194}

In many cases, chemical changes and mutations have little effect on protein stability. This is one reason that mutagenesis techniques are commonly employed to investigate protein function. In other cases, however, modification sufficiently alters the thermodynamic stability of the protein to encourage aggregation. Alterations to the surface are tolerated better by most proteins than those occurring in the core. This is probably because changes in the core synergistically affect packing efficiency by disrupting hydrophobic interactions that drive folding or remove hydrogen bonds within the core that stabilize polar or charged species. On the surface, changes that decrease stability may exert their effect by diminishing the net charge of the protein, disrupting specific electrostatic networks or increasing exposure of the apolar surface area to promote aggregation. It is difficult to predict the effect of a particular structural change on protein association because the site at which chemical modification occurs may not directly participate in the aggregation-inducing
protein–protein interaction but instead may alter the energy landscape such that conformational dynamics allow the protein to more frequently sample less condensed states. This increases the probability that stable nonnative associations will result from intermolecular collisions.

## 1.7 Approaches to Predicting Aggregation

Methods for predicting protein aggregation are based on fundamental physicochemical properties of the heteropolymer. The two most common approaches encompass evaluation of (1) the chemical composition or amino acid sequence and (2) structural features based on compatibility with the known fibrillar cross-β structure or propensity to form β-structure in the native state. Within each framework, several variations have been developed that have differing degrees of complexity, accuracy, and validity with respect to experimental data.

A sequence-based approach to predicting aggregation from the properties of individual amino acids was developed by Dobson and coworkers. It is based on the propensities of each residue to participate in aggregation and to adopt a beta structure.\(^{125}\) The propensity to aggregate depends on the hydrophobicity of an individual side chain and also on the pattern of hydrophobic residues in which the residue is embedded. The propensity to form a beta structure is countered by the inclusion of a term for α-helical propensity. The net charge is considered and may be modulated to account for the effect of pH. An overall Z-score is assigned so that peptides of the same length with different sequences may be ranked relative to a randomly generated sequence. When the algorithm was applied to Aβ(1-42), α-synuclein, and tau, the highest Z-scores for aggregation corresponded to the regions of the sequence known to form amyloids, confirming that these fundamental properties of proteins correlate well with aggregation propensity.

The Chiti–Dobson equation was further developed to predict the rate of aggregation, specifically elongation or growth of the aggregate following nucleation, for mutated proteins. The assessment is based on the protein composition and compares the relative stability of individual mutants to the wild-type protein using first principles derived from protein folding and aggregation experiments. The equation includes terms for hydrophobicity, propensity to form α-helix and β-strand structure, and net charge. The equation used to calculate aggregation rate is

\[
\ln \left( \frac{v_{\text{mut}}}{v_{\text{wt}}} \right) = 0.633(\Delta \text{Hydr}) + 0.198(\Delta \Delta G_{\text{coil-α}} + \Delta \Delta G_{\text{β-coil}}) - 0.491(\Delta \text{charge})
\]

where the change in hydrophobicity is \(\Delta \text{Hydr}\), the difference in free energy between the random coil and structured state is \(\Delta \Delta G_{\text{coil-α}}\) for alpha and \(\Delta \Delta G_{\text{β-coil}}\) for beta structure, and the net change in charge is \(\Delta \text{charge}\). Instability is defined as the inverse of the normalized \(\Delta \Delta G\) or Tm. Several comparisons to
epidemiological data sets have been performed, and the results validate this computational approach. On average, ΔΔG or Tm values correlate very well with the sum of the predicted instability and aggregation propensity, with instability being a better predictor of disease onset than aggregation propensity when each is considered independently.\textsuperscript{195}

The Zyggregator method presented by Tartaglia and Vendruscolo also predicts protein aggregation propensity from the amino acid sequence using the same first principles as Dobson–Chiti, but this algorithm includes terms describing the pattern or spatial relationship between residues and takes gatekeeper residues into account.\textsuperscript{196} The charge is also considered in a more distributed manner and is determined using a sliding window of 21 residues. Analyses of Aβ(1-42) and α-synuclein were performed, and the computational results obtained using only intrinsic factors correlate well with experimental data, identifying specific segments of the sequence responsible for aggregate formation. Moreover, when a competing term for the intrinsic propensity of the protein to form the native structure is included based on the CamP method, the approach yields higher accuracy results, enabling evaluation of different modes of aggregation within the same protein sequence. Zyggregator further includes linear correction terms to account for the extrinsic factors pH and ionic strength of the solution. The method was applied using a seven-residue window size to several additional peptides and proteins known to aggregate to more thoroughly investigate the relationship between structural protection and aggregation propensity.\textsuperscript{197}

A model to predict aggregation-prone sequences from fundamental properties of the polypeptide chain and two environmental factors was also developed by Cafisch and coworkers.\textsuperscript{198} The approach also predicts the orientation of the strands in the sheet as parallel or antiparallel. A genetic algorithm is used to optimize sequence space and conformational sampling is accomplished with MD. The computation yields the aggregation propensity along the sequence and the rate of aggregation. The absolute rate equation includes terms that account for aromaticity, polarity, solubility, and formal charges in addition to the terms used in the other computations described above. Orientation of the strands is primarily a reflection of the electric dipole moment of the peptide, such that polar residues favor an antiparallel arrangement due to the presence of unfavorable dipole–dipole interactions in an in-register parallel sheet organization. Concentration and temperature are included in the determination of the rate. This method achieves a high degree of accuracy in predicting both rate and structure.

TANGO is a statistical mechanical model designed to predict β-aggregation in peptides and proteins as a means of identifying nucleation sites for aggregation.\textsuperscript{199} The approach is based on the observation that aggregates often contain an increased β-structure and the core regions of an aggregate are completely buried, such that nucleating sequences will have their hydrogen-bonding potentials largely satisfied. The algorithm takes into account competing conformations, including α-helix, beta sheet, the folded state, and β-aggregates,
to calculate the partition function of the conformational phase space. The frequency at which a conformation is sampled depends on an energy function derived from statistical and empirical factors. Because packing interactions are considered, the approach differentiates between β-propensity and β-aggregation. TANGO accurately predicted aggregation or no aggregation for the majority of a set of 179 real peptides reported in the literature. Twenty-one false-positive and three false-negative results were found, and it was determined that peptides with low aggregation propensities were not identified using this method. The program can be interfaced after registering via the Web at http://tango.embl.de/ and is free for academic use.

Because many aggregates have been found to correspond to short peptide fragments derived from larger proteins, a library of 20-residue peptides was generated from which aggregation-prone elements were identified. Based on the results, a computational approach known as peptide interaction matrix analyzer (PIMA) was developed to predict aggregation of proteins from sequences containing β-prone elements. The protein sequence is divided into partially overlapping fragments of a specific length, and each peptide is threaded onto an in-register parallel or antiparallel beta-sheet structure. The energy associated with interaction between elements is calculated and used to predict aggregation tendencies. The study revealed that the prediction is best when 10-mer segments are used. Seven-mer fragments produce energies too small to effectively discriminate among and rank aggregation-prone sequences, whereas predictions were less accurate when based on 20mers, presumably because not all 20 residues participated in the association. When peptides derived from amyloid-forming proteins were examined, the results matched experiments well.

A web-based server for prediction of amyloid structure aggregation (PASTA) was developed based on the in-register parallel alignment observed in most cross-β structures. The assessment is made in two dimensions and involves determining first the probability of aggregation on a per-residue basis followed by the probability of cross-strand pairing with self-alignment. Cross-strand association is determined using a pairwise energy function to evaluate facing residues in the sheet. As a result, PASTA also provides an assessment of intermolecular hydrogen bond pairings between strands. The result is output as a contact map highlighting the probability of amyloid behavior. The freely accessible site is housed at http://protein.cribi.unipd.it/pasta/ and the source code is available for download.

AGGREGSCAN is a sequence-based aggregation prediction tool available on the Web. The method relies on an aggregation propensity scale for each of the naturally occurring amino acids and is applied to short segments of the protein sequence to identify hot spots for aggregation. The window size can be varied and has relatively little impact on the result using this approach, except at the extreme limits where very short sequences (less than five residues) generate numerous hot spots, many of which do not correspond to experimental findings, and with very long stretches relative to the length of
the polypeptide being examined. The method was trained against 57 amyloidogenic proteins for which the aggregation-prone sequences were experimentally determined. It can be applied to proteins having up to 2000 residues, and multiple sequences can be input together for comparative analysis. Results are returned in minutes via the Web at http://bioinf.uab.es/aggrescan/.

The Dokholyan lab has developed the iFold server to enable large-scale simulations of folding and thermal unfolding using structure-based, coarse-grained models and DMD. While the model is not specifically designed for analyzing aggregation, unfolding is a precursor to assembly, and assessing modes of unfolding for a protein may be useful in the determination of the accessibility of regions involved in aggregate formation. The primary goal of iFold is to investigate dynamics, and as some aggregate structures are known, their dynamics may be assessed. The server is found at http://iFold.dokhlab.org/. The same group developed H-predictor to identify hinge regions in proteins that are responsible for domain swapping. A two-bead model estimates the enthalpy change upon hinge formation based on alterations in bonds and native contact distances. The method was applied to predict domain-swapped dimers of RNase A and focal adhesion targeting (FAT) domain proteins. H-predictor is available at http://dokhlab.unc.edu/tools/h-predictor.

Analysis tools have been developed to identify intrinsically disordered regions in proteins based on the amino acid sequence. Often, intrinsically disordered proteins resist aggregation and their sequences contain high percentages of charged residues. PONDR and DISEMBL are web-based approaches for predicting intrinsic disorder. Disordered sequences may also be predicted to have a $\beta$-structure, but they are not prone to self-association because, similar to globular proteins, they too have endured evolutionary selection to avoid aggregation in vivo. This complementary analysis provides a means of identifying sequences within proteins that are unlikely to contribute to aggregation when unfolded.

1.8 CONCLUSIONS

A great increase in our understanding of protein aggregation has been gained as a result of comparisons to the more thoroughly studied and developed field of protein folding and unfolding. The models created to evaluate the importance of individual factors, both intrinsic and extrinsic, on kinetic and thermodynamic aspects of protein unfolding have been modified to investigate aggregating peptides and proteins and the mechanisms responsible for nucleation and elongation in aggregate formation. Examination of insoluble fibers using medium- to high-resolution experiments has provided molecular details about the patterns observed in regular aggregates as well as some specific interatomic distances between side chains. These details, along with the recognition that aggregates are often organized into beta sheets, have enabled in silico approaches that are beginning to elucidate the fundamental principles
AGGREGATION OF THERAPEUTIC PROTEINS

underlying assembly. So far, the majority of attention has been paid to fibrillar aggregates because they are frequently correlated with disease states and their repeating register is more amenable to analysis using existing experimental techniques. Computational approaches now need to be applied based on the insights gained from studying fibrillar systems, and new tools and analysis methods need to be developed to better characterize amorphous aggregates. Amorphous aggregates are more common, and their formation diminishes the efficacy and safety of protein therapeutics and also frustrates production and limits the shelf life of biotechnology products. An increased understanding of the molecular basis for protein aggregation is essential for our being able to stabilize proteins against aggregation in vitro and to develop effective treatments for protein misfolding diseases.

REFERENCES


