Potential aggregation prone regions in biotherapeutics
A survey of commercial monoclonal antibodies

Xiaoling Wang, Tapan K. Das, Satish K. Singh and Sandeep Kumar*

Pharmaceutical R & D; Global Biologics; Pfizer Global Research & Development; Chesterfield, MO USA

Abbreviations: mAbs, monoclonal antibody; FDA, food and drug administration; CDR, complementarity determining region; Fab, fragment, antigen binding; Fc, fragment, crystallize easily; Fv, fragment, variable; V_{L}, variable domain of light chain; V_{H}, variable domain of heavy chain; scFv, single chain Fv; C_{L}, constant domain of light chain; C_{H}, constant domain of heavy chain; PDB, protein data bank; Q/N, glutamine and/or asparagine; QbD, quality by design

Key words: monoclonal antibody, aggregation, antibody sequence, aggregation-prone region, aggregation prediction

Aggregation of a biotherapeutic is of significant concern and judicious process and formulation development is required to minimize aggregate levels in the final product. Aggregation of a protein in solution is driven by intrinsic and extrinsic factors. In this work we have focused on aggregation as an intrinsic property of the molecule. We have studied the sequences and Fab structures of commercial and non-commercial antibody sequences for their vulnerability towards aggregation by using sequence based computational tools to identify potential aggregation-prone motifs or regions. The mAbs in our dataset contain 2 to 8 aggregation-prone motifs per heavy and light chain pair. Some of these motifs are located in variable domains, primarily in CDRs. Most aggregation-prone motifs are rich in β branched aliphatic and aromatic residues. Hydroxyl-containing Ser/Thr residues are also found in several aggregation-prone motifs while charged residues are rare. The motifs found in light chain CDR3 are glutamine (Q)/asparagine (N) rich. These motifs are similar to the reported aggregation promoting regions in prion and amyloidogenic proteins that are also rich in Q/N, aliphatic and aromatic residues. The implication is that one possible mechanism for aggregation of mAbs may be through formation of cross-β structures and fibrils. Mapping on the available Fab—receptor/antigen complex structures reveals that these motifs in CDRs might also contribute significantly towards receptor/antigen binding. Our analysis identifies the opportunity and tools for simultaneous optimization of the therapeutic protein sequence for potency and specificity while reducing vulnerability towards aggregation.

Introduction

Therapeutic monoclonal antibodies (mAb) are a class of medications derived from living organism and produced by means of biologic techniques such as hybridoma, recombinant DNA and phage display. Monoclonal antibodies are highly specific, have high affinity as well as selectivity for binding with therapeutic targets and exhibit less degree of non-mechanism toxicity. The antibodies can also be exogenously engineered in commercially viable manner. The fraction of therapeutic mAbs has been increasing in the research and development portfolios of pharmaceutical industry. So far, more than 20 therapeutic mAbs with various indications have been approved by US Food and Drug Administration (FDA) and hundreds of monoclonal antibodies are currently in various stage of development.

Antibodies have been studied for over a century, so the basic structures and functions of antibodies are now well understood. Briefly, the building units of antibodies consist of two identical light and two identical heavy polypeptide chains held together by disulfide bridges. Light chains have two isotypes, namely λ and κ, differing in sequence composition. Heavy chains have five isotypes based on chain structure and effector function. All currently approved therapeutic mAbs belong to the IgG class. IgG has the simplest form and is the major immunoglobulin type in human sera. IgGs are further divided into four subclasses, IgG1, IgG2, IgG3 and IgG4. The sequences of light chains and different IgG heavy chains are variable in the N-terminal variable domain and conserved in the remaining constant domain. The variable domains, especially the complementarity determining regions (CDRs), determine the antigen binding specificity. Each IgG subclass has characteristic disulfide bond pattern, differing mostly in the hinge region. The hinge is the most flexible region in a mAb. It connects the two Fab and the Fc domains into ‘Y’ or ‘T’ like overall structures as revealed from the crystal structures for intact murine antibodies.

Monoclonal antibodies, like other protein therapeutics, are susceptible to degradation at all stages from production to dose...
aggregation. Several chemical and physical pathways, such as deamidation, oxidation, hydrolysis/fragmentation, isomerization and aggregation, are responsible for the degradation of the proteins. Among these, aggregation is the least understood degradation route for mAbs.

Aggregation is widely seen in proteins and refers to self-association of a number of protein molecules that may form visible or invisible particles, precipitates or fibrils. Self-association may or may not be preceded by adoption of non-native conformations by the protein or parts thereof. Aggregation is an important factor in the development of a biotherapeutic. The primary concern around aggregation is the potential to enhance immune response, which presents a risk-factor in the development of these therapeutics. If aggregation in the liquid state can not be controlled, the product may need to be lyophilized, increasing the cost of production. Aggregation problems are further exacerbated when the mAb needs to be formulated at high concentrations for subcutaneous route of administration. Hence, scientists engaged in product development and formulation of the biologic drug products must discover ways to counter these problems. In the absence of a molecular-level understanding of the aggregation mechanism, work in preventing aggregation has been mainly based on experimental heuristic screening, and dependent on particulars of each case.

Protein misfolding and aggregation is an intrinsic property that underlies a variety of neurodegenerative disorders such as Alzheimer and Parkinson disease. A hallmark of these diseases is aggregation by formation of fibrillar amyloid deposits. Through extensive research, sequence motifs that lead to this pathogenesis have been identified. Computational tools have been developed to analyze protein sequences and predict the propensity for such aggregate formation. Two such sequence-based prediction tools are TANGO and PAGE.

We believe that the intrinsic (in)stabilities of mAbs, as governed by the sequence-structure-function paradigm, are also coded in their sequences. In this report, we have collected the sequences of commercial mAbs and performed sequence analysis in a systematic manner. We used sequence-based prediction tools (TANGO and PAGE) to identify the potential aggregation-prone regions in commercial mAb sequences. Our analyses reveal several common features among the aggregation-prone motifs found in both variable and constant domains. The aggregation-prone motifs in variable domains are primarily located in the CDR loops, the regions that bind directly with the cognate receptor or antigen. Almost all aggregation-prone motifs are rich in hydrophobic or aromatic residues. Hydroxyl group containing residues, Ser and Thr, are also found in several aggregation-prone motifs. However, charged residues (Asp, Glu, Lys, Arg and His) are rare in such regions. The aggregation-prone motifs found in light chain CDR3 regions are glutamine (Q)/asparagine (N) rich. Sequence motifs rich in Q/N, hydrophobic or aromatic residues have been previously implicated in aggregation of prion or amyloidal proteins.

We also performed identical analyses on a set of non-commercial antibody sequences and obtained similar results. The importance of location of aggregation-prone regions on 3D structure is also addressed. Mapping of the aggregation-prone motifs on the available Fab—receptor/antigen complex structures revealed that, when present in mAb CDRs, these motifs might also contribute towards receptor/antigen binding. This poses an interesting challenge to antibody engineers: how to minimize the vulnerability of the molecule towards physical (or chemical) degradation while still retaining high potency and specificity towards the cognate targets?

We must also point out that our analysis delves into only one part of the complicated phenomena of aggregation. Apart from the intrinsic factors, there are a whole host of extrinsic factors that together ultimately determine the levels of aggregation seen in a product. Furthermore, judicious process and formulation development diminishes the impact of intrinsic factors and attenuates the impact of extrinsic factors, as evidenced by the number of successful commercial mAb products that have had a significant impact on human health over the last two decades.

**Therapeutic Monoclonal Antibodies in the Market**

Since the launch of muromonab-CD3 (Orthoclone OKT3) in 1986, therapeutic monoclonal antibodies have evolved from the early murine origin, through chimeric and humanized, to fully human form in little over 20 years. All 22 FDA approved commercial therapeutic monoclonal antibodies, as well as two polyclonal antibody fragments and a diagnostic mAb approved by FDA, are listed in Table 1. Indications, formulations and marketing for these antibodies are summarized in refs. 29–31. The increase in the number of therapeutic antibody drug products over time mirrors the advances in antibody production techniques. The first therapeutic candidates were murine antibodies produced by hybridoma technique, but use of these mAbs was limited due to their immunogenic complications in human patients. Patients produced human anti-mouse antibody (HAMA) response as the murine antibody was 100% foreign protein. Due to these concerns, production of murine mAbs acrituomab (CEA-scan, www.immunomedics.com), fanolesomab (Neutrospec, www.palatin.com), imicromab pentetate (Myocinsc, www.fda.gov/cder), satumomab pendetide (Oncoscint, http://nuclearpharmacy.uams.edu/resources/PackageInserts.asp), and nofetumomab (Verluma, http://nuclearpharmacy.uams.edu/resources/PackageInserts.asp), was discontinued.

Subsequent generations of therapeutic antibodies have been re-engineered to reduce mouse origin content in order to reduce immunogenicity. In the mid-1980s, simple chimeric mouse-human antibodies were made using recombinant DNA techniques. Chimeric antibodies consist of variable domain of mouse antibody and constant domains of human antibody. The mouse content is reduced to approximately 33% in chimeric mAbs. Two chimeric mAbs, rituximab (Rituxan) and infliximab (Remicade) are among the five top selling antibodies today. Antibodies were further humanized by grafting only murine CDR regions into the human antibody framework. Humanized antibodies have 5–10% of mouse content. Following the approval of the first humanized therapeutic antibody (daclizumab [Zenapax]) in 1997, humanized antibodies trastuzumab (Herceptin) and bevacizumab (Avastin) have gone on to achieve blockbuster status.
Most commercial antibodies are intact conventional mAbs. However, for enhanced affinity, reduced toxicity and improved tissue permeation, antibody fragments such as Fab, Fv and single chain Fab (scFv) can be used. The Fab products are currently most common. Marketed Fabs (Table 1) are ranibizumab (Lucentis),

Table 1  List of commercial mAbs

| # | Brand name  | Molecule name         | Isotype and format | Source       | Source of sequence | Approval year | Company                      |
|---|-------------|-----------------------|--------------------|--------------|--------------------|---------------|------------------------------|
| 1 | Avastin     | Bevacizumab           | IgG1 κ             | Humanized    | CAS ID: 216974-75-3| 2004          | Genentech and BioOncology   |
| 2 | Bexxar      | Tositumomab and I-131Tositumab | IgG2a λ; Conjugated to Iodine | Murine       | DB ID: DB00087     | 2003          | Corixa and GSK              |
| 3 | Campath     | Alemtuzumab           | IgG1 κ             | Humanized    | DB ID: DB00087     | 2001          | Ilex Oncology; Millenium and Berlex |
| 4 | Cimzia      | Cetolizumab pegol     | IgG1 K; Fab; Conjugated to Polyethylene glycol | Humanized    | CAS ID: 428863-50-7 | 2008          | Nektar/UCB                   |
| 5 | CroFab      | Croataliidae Polyvalent Immune Fab | Fab               | Ovine        | DB ID: DB00076 PDB ID: 1IGJ | 2002          | Protherics Inc              |
| 6 | DigiFab     | Digoxin Immune Fab    | IgG2a δ; Fab       | Ovine        | DB ID: DB00076 PDB ID: 1IGJ | 2001          | Protherics Inc              |
| 7 | Erbitux     | Cetuximab             | IgG1 κ             | Chimeric     | DB ID: DB00002     | 2004          | ImClone and BMS             |
| 8 | Herceptin   | Trastuzumab           | IgG1 κ             | Humanized    | DB ID: DB00072     | 1998          | Genentech                   |
| 9 | Humira      | Adalimumab            | IgG1 κ             | Human        | CAS ID: 331731-18-1| 2002          | CAT and Abbott              |
| 10 | Lucentis    | Ranibizumab           | IgG1 κ; Fab        | Humanized    | CAS ID: 347396-82-1| 2006          | Genentech                   |
| 11 | Mylotarg    | Gemtuzumab ozogamicin | IgG4 κ; Fab        | Humanized    | ref. 36, Fv only   | 2000          | Celltech and Wyeth          |
| 12 | Orthoclone  | Muromonab-CD3         | IgG2a κ            | Murine       | DB ID: DB00075     | 1986          | Ortho Biotech               |
| 13 | ProstaScint | Indium-111 Capromab pendetide | IgG1 κ; Conjugated to Indium | Murine       | DB ID: DB00075     | 1996          | Cytogen                     |
| 14 | Raptiva     | Efalizumab            | IgG1 κ             | Humanized    | CAS ID: 214745-43-4 ref. 37 | 2003          | Genentech and Xoma          |
| 15 | Remicade    | Infliximab            | IgG1 κ             | Chimeric     | ref. 38            | 1998          | Centocor                    |
| 16 | ReoPro      | Abciximab             | IgG1 κ; Fab        | Chimeric     | DB ID: DB00073     | 1997          | Centocor/Lilly              |
| 17 | Rituxan     | Rituximab             | IgG1 κ             | Chimeric     | DB ID: DB00074     | 1998          | Genentech and IDEC          |
| 18 | Simulect    | Basiliximab           | IgG1 κ             | Chimeric     | DB ID: DB00074     | 1998          | Novartis                    |
| 19 | Soliris     | Eculizumab            | IgG2/4 κ           | Humanized    | CAS ID: 219685-50-5 ref. 39, 40 | 2007          | Alexion Pharmaceuticals      |
| 20 | Synagis     | Palivizumab           | IgG1 κ             | Humanized    | PDB ID: 2HWZ ref. 41 | 1998          | Medimmune                   |
| 21 | Tysabri     | Natalizumab           | IgG4 κ             | Humanized    | CAS ID: 189261-10-7 | 2004          | Elan and Biogen Idec       |
| 22 | Vectibix    | Panitumumab           | IgG2 κ             | Human        | CAS ID: 339177-26-3 | 2006          | Amgen                        |
| 23 | Xolair      | Omalizumab            | IgG1 κ             | Humanized    | CAS ID: 242138-07-4 | 2003          | Genentech w Novartis        |
| 24 | Zenapax     | Daclizumab            | IgG1 κ             | Humanized    | CAS ID: 152923-56-3 | 1997          | Roche                        |
| 25 | Zevalin     | Ibritumomab           | IgG1 κ; Conjugated to Indium or Yttrium | Murine       | DB ID: DB00078     | 2002          | IDEC                         |

The current mAb market is dominated by chimeric and humanized antibodies. However, fully human antibodies generated using phage display technique and transgenic animals are reaching the market. Adalimumab (Humira; launched in 2002) is one such example.
abciximab (ReoPro), certolizumab (Cimzia), digoxin immune Fab (DigiFab) and crotalidae polyvalent immune Fab (CroFab). The latest technologies for the design of high affinity antibodies is reviewed by Wark et al.52

**Sequence Differences and Similarities**

**Sequence availability for commercial mAbs.** We have collected the sequences of commercial mAbs in as complete a manner as feasible to perform sequence analysis. The sources of sequences are Drug Bank database (www.drugbank.ca)53 and CAS registry (www.cas.org). Some partial sequences are collected from RSCB Protein Data Bank (www.rcsb.org/pdb/home/)54 and literature. The Drug Bank ID, CAS ID, PDB ID, and references are listed in Table 1.

The sequences of commercial mAbs are not all publicly available. For example, we could not find the sequences of tositumomab (Bexxar), abciximab (ReoPro), capromab pendetide (ProstaScint) and crotalidae polyvalent immune Fab (CroFab) in the public domain. In a few other cases, only partial sequences are publicly available. The sequences of only the Fab and Fv regions are available for palivizumab (Synagis) and gemtuzumab ozogamicin (Mylotarg), respectively. Fv sequence of infliximab (Remicade) has been published,58 but the CDR regions are missing. Thus, infliximab (Remicade) was not included in subsequent sequence analysis.

**Sequence isotypes of the commercial mAbs.** Table 1 summarizes the brand name, generic name, isotype, source, year of first FDA approval, and the manufacturer for each antibody included in this study. The heavy chains are dominantly IgG1 with only a few of the IgG2 or IgG4 type. The choice of IgG isotype commonly depends on the requirement for complement binding and effector function.55,56 IgG1 has a flexible hinge and good Fc-mediated effector activity, while IgG2 has limited effector activity. IgG2 is commonly chosen when only the binding specificity is desired.56,57 However, mAbs of IgG2 type have been reported to show disulfide scrambling.58-60 Disulfide scrambling affects the hinge structure in IgG2 mAbs and likely leads to product heterogeneity.

Most light chains are of κ type. Tositumomab (Bexxar) is the only product with λ light chain. The light chain type for abciximab (ReoPro) and crotalidae polyvalent immune Fab (CroFab) is unknown from public sources. κ light chains have a high abundance in mice (95%) but much lower (60%) in humans.61

**Sequence alignments for commercial mAbs.** The multiple sequence alignment of light chains for commercial mAbs is presented in Figure 1A. All the light chains shown in Figure 1A are κ chains. The degree of sequence identity is consistent with the sequence origin, i.e., source organism. The sequences of human and humanized mAbs have very high identities. The pair-wise aligned scores are as high as 90%. As expected, the differences primarily lie in CDR regions. The aligned scores are reduced to 80% if only V L sequences are considered. The aligned scores between the human or humanized and the chimeric mAbs are 75–80%. Even though ibritumomab (Zevalin) is a murine mAb, its light chain is very similar to the two chimeric mAbs, basiliximab (Simulect) and rituximab (Rituxan). Their aligned scores are close to 90%. The murine mAb muromonab-CD3 (Orthoclone OKT) and the ovine mAb digoxin immune Fab (DigiFab) share lowest sequence similarities with other mAbs, as indicated by the low aligned scores of ~60%.

Figure 2A shows the heavy chain sequence alignment for commercial mAbs. Similar to light chains, sequences in constant regions for mAbs of the same isotype are very similar, if not identical. Variability comes from V H especially CDRs. In the IgG1 group, overall sequence identities of human and humanized mAbs (adalimumab [Humira], bevacizumab [Avastin], alemtuzumab [Campath], trastuzumab [Herceptin], efalizumab [Raptiva], omalizumab [Xolair], dactizumab [Zinapax], certolizumab [Cimzia Fab] and ranibizumab [Lucentis Fab]) are greater than 85%, while the V H sequence aligned scores are ~67%. The same applies to the chimeric IgG1 mAbs cetuximab (Erbitux), basiliximab (Simulect) and rituximab (Rituxan). These mAbs share 85% overall sequence identities, but 54% identities for V H sequence only. The humanized mAb palivizumab (Synagis) appears to be an exception as its aligned scores with other human or humanized IgG1 mAbs are below 70% when only Fab portions of the sequences are aligned (note that only Fab sequence is available for palivizumab). Ibritumomab (Zevalin) is the only IgG1 murine mAb in Figure 2A. Its sequence similarities with others are lower than 60%.

The hinge region is important for flexibility and effector functions of the IgG molecules.55 Each IgG subclass has characteristic sequence composition and disulfide bonding pattern in the hinge region.62,63 All the IgG1 commercial mAbs, except ibritumomab (Zevalin), have the characteristic IgG1 hinge with three cysteine residues (Fig. 2A). The first cysteine is disulfide bonded with the C-terminal cysteine in light chain. The other two cysteines form inter-heavy chain disulfide bridges. The constant region of eculizumab (Soliris) is the hybrid of IgG2 (C H2 and hinge region) and IgG4 (C H3 domain) and IgG4 (C H2 and C H3 domain).64 Both eculizumab (Soliris) and panitumumab (Vectibix) have the typical IgG2 hinge. However, the hinge sequence for panitumumab as published in CAS registry (CAS ID: 339177-26-3) does not show an IgG2 characteristic pattern. In Figure 2A, the sequence of panitumumab (Vectibix) is shown after replacement with a canonical IgG2 hinge. Muromonab-CD3 (Orthoclone OKT) is another outlier as an IgG2 mAb with an IgG1 hinge. Natalizumab (Tysabri) has the IgG4 hinge in which the four cysteines form inter-heavy chain disulfide bonds.

**Sequence alignments for non-commercial antibodies.** To compare our data set of commercial mAbs, we have randomly collected 20 non-commercial antibody sequences. The crystal structures for all of these are available in the protein data bank (PDB).54 Three of the 20 are the sequences for intact mAbs (PDB entries: 1IGT, 1IGY, 1HZH), seven are antigen-antibody (Fab) complexes and the remaining ten are FabS only. We have done the same analyses on these 20 antibody sequences. Their light and heavy chain sequence alignments are presented as Figures 1B and 1B to facilitate the direct comparison with the commercial mAbs.
Figure 1. For figure legend, see page 259.
Aggregation prone regions in commercial mAbs

Figure 1. (A) Sequence alignment of light chains of commercial mAbs. The letters at the end of the mAb brand name are used to indicate its source. ‘u’ for human, ‘zu’ for humanized, ‘xi’ for chimeric, ‘o’ for mouse and ‘ov’ for sheep mAb. For molecular name of each mAb, refer to Table 1. All cysteine amino acids are highlighted in green. The three CDRs are highlighted in yellow. The predicted aggregation-prone segments are in red letters.

(B) Sequence alignment of light chains of the twenty randomly selected non-commercial mAbs. Highlighting has the same meaning as in (A).

Figure 2A. For figure legend, see page 260.
Aggregation prone regions in commercial mAbs

Commercial mAbs are subjected to numerous stress factors, such as pH changes, temperature excursions, chemical modification, shear and freeze/thaw, at various stages of production, shipping and storage. These stress factors may trigger self-association of the mAb molecules leading to aggregation. Currently, the process by which protein molecules self-associate to seed aggregation and the total number of variables involved are not well-understood. Does molecular association require a conformational change from the native state? If so, what degree of conformational change is required? Which regions of the molecule are more conformationally labile than others? Are these conformationally labile regions more likely to drive self-association and promote aggregation? Accumulating experimental data shows that proteins, especially ones with amyloidogenic properties, often contain specific sequence or structural motifs that tend to promote aggregation. Knowledge of the aggregation-prone regions in proteins may help alleviate aggregation concerns in biotherapeutics via rational design and/or selection of robust yet potent candidates, thus, bringing Quality by Design (QbD) approaches to the early stages of drug discovery and development.

Prediction of potential aggregation-prone regions

Commercial mAbs are subjected to numerous stress factors, such as pH changes, temperature excursions, chemical modification, shear and freeze/thaw, at various stages of production, shipping and storage. These stress factors may trigger self-association of the mAb molecules leading to aggregation. Currently, the process by which protein molecules self-associate to seed aggregation and the total number of variables involved are not well-understood. Does molecular association require a conformational change from the native state? If so, what degree of conformational change is required? Which regions of the molecule are more conformationally labile than others? Are these conformationally labile regions more likely to drive self-association and promote aggregation? Accumulating experimental data shows that proteins, especially ones with amyloidogenic properties, often contain specific sequence or structural motifs that tend to promote aggregation. Knowledge of the aggregation-prone regions in proteins may help alleviate aggregation concerns in biotherapeutics via rational design and/or selection of robust yet potent candidates, thus, bringing Quality by Design (QbD) approaches to the early stages of drug discovery and development.

Several molecular seeding conformational changes may give rise to aggregates. One of the best-studied conformational seed is the cross-β motif seen in amyloid-like fibrils. While the presence of β-strands is not essential to form this motif, several β-rich proteins (e.g., β2-microglobulin, transthyretin, immunoglobulin light chain especially of type λ, immunoglobulin heavy chain, concavalin A) form amyloids using the cross-β motif as seed. In addition, Siepen et al. have shown that the exposed edge β-strands in...
proteins can initiate amyloidogenesis by forming the cross β-motifs. For example, the recognition between exposed edge β-stands is an important step in amyloid fibrilogenis for transthyretin and β2-microglobulin. In another example, Pande et al. has proposed that aggregation by P23T mutant of human γD-Crystallin may be via this mechansim. Moreover, work from Dobson's group has shown that fibrillogenesis is a general feature of proteins under stress or denaturing conditions. Monoclonal antibodies are also β-sheet rich proteins. Hence, we were curious to explore if mAbs can also potentially aggregate via the cross β-motif conformational seed. If so, the sequence motifs prone to this type of aggregation should be found in both the commercial and other mAbs. The paper by Maas et al. is particularly relevant in this regard. Using the amyloid-marker dyes Congo Red and Thioflavin T, they detected formation of cross-β structures and amyloid-like fibrils in many biopharmaceuticals, including two antibodies which nevertheless had among the lowest signal levels. Besides this, we also have evidence for Thioflavin T dye binding in aggregated mAb samples from our own work in progress. This work is our first attempt at identifying aggregation-prone motifs in therapeutic mAb sequences by taking advantage of the tools developed for amyloidogenic proteins. The presence of just one aggregation-prone motif could drive cross-β motif type aggregation. However, the occurrence of aggregation-prone sequences in mAbs does not guarantee that aggregation shall necessarily proceed via this mechanism. This may or may not be the significant mode of aggregation in therapeutic mAbs.

**Prediction methods.** A number of computational approaches, such as TANGO, PAGE, and Zyggregator have been developed to predict the aggregation-prone regions in amyloidogenic proteins. We have chosen two prediction tools, viz. TANGO and PAGE, to identify potential aggregation-prone regions in commercial mAb sequences. TANGO and PAGE have been parameterized based on the experimental data on amyloidogenic peptides and proteins. However, the two tools use different algorithms.

TANGO and PAGE, developed by Serrano's and Caflisch's group respectively, use the primary amino acid sequence only and make predictions based on the physicochemical properties of amino acids. The TANGO algorithm relies on physicochemical rules behind β-sheet formation and assumes that the core of an aggregate is completely desolvated. TANGO calculates relative probabilities of peptide segments to exist in different conformational states such as α-helix, β-strand, turn, random coil and β-aggregate. The TANGO algorithm was tested against 179 peptides derived from 21 different proteins as published. TANGO yielded a success rate of 92% for aggregation tendency of 5% or greater.

PAGE computes the aggregation propensity (lnR) and absolute aggregation rate (lnR) for a given amino acid sequence by sliding a small window of 5–9 residues along the sequence. The aggregation propensity is calculated based on aromaticity, β-strand propensity, charge, solubility, average polar/non-polar/accessible surface area of each residue in the given widow. PAGE was tested against peptides from 16 proteins including α-synuclein, apolipoprotein, APP, IAPP, prion, Sup35 and tau. We deliberately chose to use at least two different aggregation-prone region prediction tools. This approach has helped us avoid potential biases due to the training sets, parameterization and peculiarities of methods in a given tool. Even though the authors of TANGO suggest a region is potentially aggregation-prone if the TANGO scores of more than five consecutive residues are above 5%, we use a much more stringent criteria of ≥50%. The aggregation propensity (lnp) from PAGE is converted to Z score in order to identify the regions with statistically high aggregation propensity. The Z score of residue i is calculated as follows,

\[ Z_i = \frac{\ln(p_i) - \ln(\bar{p})}{\text{std}(\ln(\bar{p}))} \]

where, \( \ln(p_i) \) is the average aggregation propensity of the sequence, \( \text{std}(\ln(\bar{p})) \) is the standard deviation. A region with \( Z \) score >1.96 is considered as aggregation-prone. We identify a region as aggregation-prone if it is strongly predicted by at least one program. Use of stringent criteria ensures that our predicted regions have a greater probability of being truly aggregation-prone. However, we must note that our choice is rather arbitrary, and additional aggregation-prone motifs could have been identified with less stringent criteria. The real test lies in experimental confirmation. Experiments on polypeptides containing the predicted aggregation-prone regions are planned.

We have performed TANGO and PAGE analysis on all the collected commercial mAbs and the 20 non-commercial antibody sequences. The TANGO and PAGE profiles of trastuzumab
Aggregation prone regions in commercial mAbs

| # | General location and sequence position | Aggregation-prone motifs | Occurrence in mAbs | # | General location and sequence position | Aggregation-prone motifs | Occurrence in mAbs |
|---|--------------------------------------|--------------------------|-------------------|---|--------------------------------------|--------------------------|-------------------|
| 1 | Precede and partly in CDR L2, Residue 51–60 | LLIYSASFLY LLIASAQL LLIGVAP LLIYAAV VLIFV | 7 | 9 | Right after CDR H2, Residue 72–76 | VTLMLV | 1 |
| 2 | Right before CDR L3, Residue 88–92 | LGIYF | 1 | 10 | Precede CDR H3, Residue 97–101 | IYYCV | 1 |
| 3 | Precede and mostly in CDR L3, Residue 92–98 | YCQQNNN YCQQHNE YCQQYS YCQQYN YCQQHY YCQQYD YCQQS | 8 | 11 | In CDR H3, Residue 104–118 | WYYNSNYWYF IYYFGTYF | 5 |
| 4 | Beginning of Cβ domain, Residue 119–124 | SVFIFP SVFIF TVFIF | 17 | 12 | In Cβ2, Residue 254–259 | SVFHP | 15 |
| 5 | In Cβ domain, Residue 137–141 | VVCLL VVCF | 19 | 13 | In Cβ2, Residue 263–270 | VLMISL | 1 |
| 6 | In CDR H1, Residue 27–39 | YIFSXYWQVV IFTDF YYWTWI | 3 | 14 | In Cβ2, Residue 273–279 | IVTCVV | 1 |
| 7 | Precede and partly in CDR H2, Residue 50–54 | IGAIQ IGVW IGYIS IGYI | 6 | 15 | End of Cβ2, Residue 317–324 | VVSVITVL VVSVTVV | 14 |
| 8 | In CDR H2, Residue 61–66 | TTEYN TNYNQ | 3 | 16 | In Cβ3, Residue 417–423 | GSFFLYS GSFFLYS | 14 |

*Continuous numbering in the aligned sequences as shown in Figures 1 and 2. is used for residue position. The column records the occurrences of the corresponding motifs in the mAbs in Figures 1 and 2.*
surprising because the goal of humanization is to reduce immunogenicity due to foreign origin.

In the antibody structure, CDRs form binding sites and are exposed on the surface of the Fab globules to facilitate antigen/receptor/hapten recognition. In absence of the cognate partners, the CDRs may promote self-assembly due to their conformational flexibility and solvent accessibility. However, it is not yet clear to us if the mAb self-association involves Fab:Fab or Fab:Fc association. In theory, both of these association modes are possible.

In general, the aggregation-prone motifs avoid charged residues (Asp, Glu, Lys, Arg and His). This feature is consistent with the fact that charged residues contribute favorably towards protein solubility. The net charge of an unfolded polypeptide chain is known to affect the aggregation-propensity. Surprisingly, the hydrophilic hydroxyl-containing Ser/Thr residues are found in several motifs (Table 2), probably due to their high propensity to form β-strand and low propensity to form α-helix. More interestingly, almost all aggregation-prone motifs are rich in aromatic and hydrophobic β branched residues. The aggregation-prone motifs found in several light chain CDR regions are rich in Gln and Asn residues. Those motifs are worth further discussion as they have also been implicated in driving aggregation of proteins involved in neurodegenerative diseases.

The aggregation-prone motifs in the third CDR of light chains (YCQQNNN in cetuximab [Erbitux], YCQQHNE in efalizumab [Raptiva]; motif 3 in Table 2) are rich in Gln and Asn. These motifs are similar to the aggregation-seeding motifs YQQYN and YQQQY found in the well-known yeast prion β2-Microglobulin readily form fibrils in vitro. The motifs 6 (YCQQQNNN and YQQQY) in β2-Microglobulin and 11 (VVYYSYYWF and IFYYGTTFYF in Heavy chain CDR 3) are rich in aromatic residues (Table 2). The role of aromatic residues in promoting aggregation is through the favorable π-π stacking interactions; Phe especially is found to form significant homostacking. These interactions lead to aggregates with the specific ordered pattern seen in fibrils.

### Table 3 Summary of aggregation-prone motifs in biopharmaceuticals

| Therapeutic protein | Source of sequence | Aggregation-prone motif |
|--------------------|--------------------|-------------------------|
| Albumin            | DB00096            | ALVUALFA 153 ELFFAK 325 VFLGMFLY |
| Somatropin         | DB00052            | ISSLILQ 101 LFVGA 161 GLLYC |
| Insulin aspart     | CAS 116094-23-6    | ALYLV 145 YVWQL 172 LIGALLV 195 FILFAVF |
| Factor VIII        | DB00025            | LFVEF 50 58 LFVIA 116 VYQVL 172 LIGALLV 195 FILFAVF |
| Epoetin Alfa       | DB00016            | GALL 65 ALYLV 78 ALVNL |
| Etanercept         | DB00005            | SVFLFP 259 VSVTL 322 VSVTL |
| Glucagon           | DB00040            | FVQWLM 22 |

*List of proteins is from Table 1 of ref. 17. Sequences are obtained from either Drugbank (www.drugbank.ca) or CAS registry. Listed are entry ID. Number before each motif is the position of the first residue of the motif.*
However, in their study of amyloid aggregation of several mutants of human muscle acylphosphatase, Bemporad et al. have suggested that the high aggregation propensity of aromatic residues is attributed to their hydrophobicity and intrinsic β-sheet forming propensity rather than aromaticity.91

Besides being the driving force for protein folding, the hydrophobic effect also plays an important role in protein aggregation.92,93 Most of the aggregation-prone motifs in commercial mAbs (Table 2) are rich in hydrophobic residues, especially β-branched aliphatic residues. Ile, Val and Leu are among the top aggregation-prone residues reported by de Groot et al.26 Commercial mAbs contain VVCLL in C4 and VVSVLTVL in C1H2 (motif 5 and 15 in Table 2). These are two representative motifs with high hydrophobicity. Stretches of hydrophobic residues, such as IAALL in Transthyretin have been shown to be aggregation promoting.71 The high aggregation propensity of hydrophobic stretches could come from two potential sources, i.e., the specific interactions mediated by the side chains of hydrophobic residues and the non-specific hydrophobicity per se.92 Kim et al. have tried to distinguish these two potential sources through systematic mutation in a study of Aβ42 peptide.92 Their findings suggest that the non-specific hydrophobicity is more important for aggregation than the specific side chain interactions.

Amyloid fibrils and β helices show similar conformations. Tsai et al.88 have performed a systematic analysis of β helices and revealed their sequence and structure characteristics. Some relevant characteristics are: (1) Charged residues except Asp disfavor the β sheet regions; (2) Asn and Phe show strong amide stacking interactions which stabilize β helices; (3) Ile and Val have the highest propensity to be in β strands. These features are also found in the aggregation-prone motifs identified in this study.

In order to promote aggregation, a region in a protein should satisfy the following criteria: (1) the amino acid residues in the region should have high intrinsic aggregation propensity; (2) the region should be conformationally unstable or flexible; and (3) the region should be exposed or become exposed upon conformational transition, to facilitate intermolecular interactions.77,94 In the case of commercial mAbs, the above mentioned motifs may well be buried in the native structure of mAb and, hence, blocked from aggregation. However, in case of conformational fluctuations, these motifs may become exposed. Therefore, the location of aggregation-prone segments on the 3-D structure and the conformational flexibility of mAb molecule are also important factors for predicting aggregation-prone regions with high accuracy. We are currently performing molecular dynamic simulation studies to further refine our predictions. The mapping of our sequence motifs on to the structures is given below.

3D Structures of Commercialized mAbs

IgG mAbs have very similar structure, which is in overall ‘Y’ or ‘T’ shape with Fab and Fc domains adopting the immunoglobulin fold. Currently, it is difficult to obtain crystal structures for intact antibodies. So far only three crystal structures of intact antibodies have been solved with moderate resolution.8,9,95 However, the Fab or Fc domains of antibodies are relatively easy to crystallize and many structures have been deposited in the Protein Data Bank. The Fab structures of some commercial mAbs are available, namely, bevacizumab (Avastin; PDB ID: 1BJ1),96 alemtuzumab (Campath; 1CE1),97 cetuximab (Erbitux; 1YY8 and 1YY9),98 trastuzumab (Herceptin; 1N8Z),99 muromonab-CD3 (Orthoclone OKT; 1SY6),100 basiliximab (Simulect; 1MIM),101 palivizumab (Synagis; 2HWZ) and digoxin immune Fab (DigiFab; 1IGJ). The Fc structure of rituximab (Rituxan; 1L6X)102 is available as well. Due to their high sequence identities, the Fab structures are strikingly similar. These Fab structures are aligned and shown in Figure 4. The root mean square distance (RMSD) values between these structures are less than 5 Å.

We have mapped the aggregation-prone regions predicted from amino acid sequence on to the 3D structure of Fabs. The aggregation-prone motifs in the CDRs are co-localized in the structures to yield characteristic structural motifs in the antigen binding site. These structural motifs may be involved in recognition of the cognate receptors or antigens. As an example, Figure 5 shows the predicted aggregation-prone regions mapped on the Fab structure of trastuzumab (Herceptin) in complex with extracellular domain of human epidermal growth factor 2 (HER2).99 HER2 plays an important role in breast cancer. Trastuzumab (Herceptin) is designed to selectively bind with high affinity to the extracellular region of HER2. Two of the four aggregation-prone segments

Figure 4. Cartoon representation of the superimposed Fab structures of bevacizumab (Avastin), alemtuzumab (Campath), cetuximab (Erbitux), trastuzumab (Herceptin), basiliximab (Simulect) and palivizumab (Synagis). All light chains are in red and heavy chains in blue.
on the Fab of trastuzumab (Herceptin) are located at CDR L2 and CDR L3, respectively. Up to 45% of the surface area of the two segments is buried after binding to HER2. This suggests that the two motifs are major participants in the binding sites. While we present trastuzumab (Herceptin) HER2 extra cellular domain as an example, our observations for the other Fab structures, except palivizumab (Synagis), are similar (data not shown).

Our observations indicate that the sequence/structural motifs in mAb CDRs that bind the cognate receptors/antigens may also be the aggregation-prone regions. If this is true, then an interesting challenge is posed to the mAb engineers: how should potent and specific mAb therapeutics with low vulnerability towards aggregation be designed? This is a significant challenge because therapeutic mAb drug products are stored in absence of their cognate partners, leaving these motifs potentially vulnerable to aggregation. Structure based biologic drug design as part of QbD approaches can help address such issues.

The other two aggregation-prone segments in trastuzumab (Herceptin) are the two closing β-strands in C_L domains of the Fabs (Fig. 5). One of the strands (SVFIFP) is at the edge of the 4-stranded β-sheet of C_L which packs against the 4-stranded β-sheet of C_H1 domain. The edge β-strands can readily form edge-to-edge aggregates. The amyloid fibril formation of a number of proteins, such as transthyretin (TTR), has been linked to the edge β-strand interactions.

**Summary and Conclusions**

The technologies driving advances in antibody development have focused on increasing the efficacy and potency of mAbs while reducing immunogenic potential. Aggregation issues with candidate mAbs are usually addressed at the formulation stage during drug development. However, consideration of aggregation issues at early stages of biologic drug discovery may save both time and resources while reducing risk at the later stages of drug formulation and clinical development. QbD approaches including structure based biologic drug design can be useful in addressing these challenges.

The aggregation-prone regions in commercial and non-commercial antibodies, predicted by TANGO and PAGE, have sequence motifs reminiscent of those found in amyloidogenic proteins. As such, this is not surprising because TANGO and PAGE have been parameterized using data from amyloidogenic peptides and proteins. However, the novel findings of our study are as follows: (1) aggregation-prone motifs in commercial and non-commercial antibodies are consistent and present at functionally important locations such as CDR loops; (2) it is possible to successfully use the methods developed for amyloidogenic peptide and proteins to predict aggregation in biotherapeutics. Sequences/motifs predicted to be involved in aggregation are seen to be important for function also. Our approach is validated by the finding that TANGO and PAGE also identified aggregation-prone motifs in non-mAb biopharmaceuticals with available experimental aggregation data (Table 3). The tools and the analyses utilized in this work can thus be applied to mutational studies to optimize activity while reducing aggregation propensity at the discovery stage. The next step in prediction of aggregation-prone regions in biologic drugs shall be incorporation of the 3D structural information.

**Acknowledgements**

We thank Drs. James Carroll, Kevin King, Sandeep Nema, Sa V. Ho, Graeme Bainbridge and Bernard Violand for several helpful discussions. The authors are grateful to Pfizer IP committee members for review of this manuscript and for several useful comments. Drs. Arindam Bose and Thomas Warren are thanked for their interest and support for our work. A postdoctoral fellowship for Xiaoling Wang via predictive stability initiative at Pharmaceutical Research and Development, Global Biologics, in Pfizer is gratefully acknowledged. Two anonymous referees are acknowledged for their critical review of the manuscript.
References

1. Kohler G, Milstein C. Continuous cultures of fused cells secreting antibody of pre-determined specificity. Nature 1975; 256:495-7.
2. Cohen SN, Chang ACY, Boyer HW, Helling RB. Construction of Biologically Functional Bacterial Plasmids in vitro. Proc Natl Acad Sci USA 1973; 70:3240-4.
3. McCarthey J, Griffiths AD, Winter G, Chiellij DJ. Phage antibodies: filamentous phage displaying antibody variable domains. Nature 1990; 348:524-2.
4. Reichert JM, Rosenweig CJ, Faden LB, Dewitz MC. Monoclonal antibody successes in clinical trials. Nat Biotechnol 2005; 23:1147-57.
5. Janeway CA, Travers P, Walport M, Sholmchik MJ. Immunobiology. Garland Science 2001. 269:120-21.

266 mAbs 2009; Vol. 1 Issue 3
68. Jones S, Manning J, Rad MD, Radford SE. Amyloid-forming Peptides from b2-Microglobulin—Insights into the Mechanism of Fibril Formation in vitro. J Mol Biol 2003; 325:249-57.
69. Bourne PC, Ramsland PA, Shan L, Fan Z-C, DeWitt CR, Shultz BB, et al. Three-dimensional structure of an immunoglobulin light-chain dimer with amyloidogenic properties. Acta Crystallogr D Biol Crystallogr 2002; 58:185-23.
70. Piskunov RR, Solomon A, Weiss DT, Stevens FJ, Schiffer M. Tertiary structure of human 16 light chains. Amyloid 1999; 6:165-71.
71. Jaroniec CP, MacPhee CE, Astefanires S, Dobson CM, Griffin RG. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. Proc Natl Acad Sci USA 2002; 99:16748-53.
72. Westermarck P. Aspects on human amyloid forms and their fibril polypeptides. FEBS J 2005; 272:592-9.
73. Verti V, Canale C, Relini A, Librizzi F, Milletello V, Gliozzi A, Leone M. Amyloid fibril formation and amorphous aggregation in concanavalin A. Biophys Chem 2007; 125:184-90.
74. Siepen JA, Radford SE, Westhead DR. B Edge strands in protein structure prediction and aggregation. Protein Sci 2003; 12:2348-59.
75. Pande A, Annunziato O, Oguo O, Benedek GB, Pande J. Decrease in Protein Solubility and Cataract Formation Caused by the Pro23 to Thr Mutation in Human gD-Crystallin. Biochemistry 2005; 44:2491-500.
76. Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G, Dobson CM. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. Proc Natl Acad Sci USA 2002; 99:16419-26.
77. Tartaglia GG, Pawar AF, Campioni S, Dobson CM, Chiti F, Vendruscolo M. Prediction of Aggregation-Prone Regions in Structured Proteins. J Mol Biol 2008; 380:42-56.
78. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D. The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci USA 2006; 103:4074-8.
79. Trejno SR, Schlez JM, Pace CN. Amino Acid Contribution to Protein Solubility: Asp, Glu and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa. J Mol Biol 2007; 366:449-60.
80. Chiti F, Dobson CM. Protein Misfolding, Functional Amyloid and Human Disease. Annu Rev Biochem 2006; 75:333-66.
81. Creighton TE. Proteins: Structures and Molecular Properties W.H. Freeman and Company 1993.
82. Pietrini EN, Pasquale LL, Gurschner MJ, Melki R, Stefani M. The Yeast Prion UrA2 Native-like Assemblies Are Toxic to Mammalian Cells Regardless of Their Aggregation State. J Biol Chem 2006; 281:15337-44.
83. Michelthch MD, Weissman JS. A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions. Proc Natl Acad Sci USA 2000; 97:11910-5.
84. Chen SM, Berthelier VB, Yang W, Wrezel R. Polyglutamine Aggregation Behavior in vitro Supports a Recruitment Mechanism of Cytoplasmic. J Mol Biol 2001; 311:173-82.
85. Yang W, Dunlap JR, Andrews RB, Wrezel R. Aggregated polyglutamine peptides delivered to nucelus are toxic to mammalian cells. Hum Mol Genet 2002; 11:2905-17.
86. Zoghbi HY, Orr HT. Glutamine repeats and neurodegeneration. Proc Natl Acad Sci USA 2000; 97:11910-5.
87. Muller YA, Chen Y, Christinger HW, Li B, Cunningham BC, Lowman HB, de Vos AM. VEGF and the Fab fragment of a humanized neutralising antibody: crystal structure of the complex at 2.4 Å resolution and mutational analysis of the interface. Structure 1998; 6:1153-67.
88. James LC, Hale G, Waldman H, Bloomer AC. 1.9 Å structure of the therapeutic antibody CAMPATH-1H Fab in complex with a synthetic peptide antigen. J Mol Biol 1999; 289:293-301.
89. Jones P, Sanzmi T, Jeffrey PJ, Weihns JF, Kussie P, Ferguson KM. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer cell 2005; 7:301-11.
90. Cho H-S, Mason K, Ramyar KX, Stanley AM, Gabelli SB, Denney DW, Lealh DJ. Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. Nature 2003; 421:756-60.
91. Kier-Nielsen LR, Dunstone MA, Kostenko L, Ely LK, Beddor D, Mifsud NA, et al. Crystal structure of the human T cell receptor CD3e/CD7 heterodimer complexed to the therapeutic mAb OKT3. Proc Natl Acad Sci USA 2004; 101:7675-80.
92. Mikol V. Structure of the Fab Fragment of SDZ CH1621: a Chimeric Antibody Against CD25. Acta Crystallogr D Biol Crystallogr 1996; 52:54-54.
93. Ishaugie EE, Peerta LG, Ganzano-Santoro H, Torpal K, Wong PY, Ultsch M, et al. Mapping of the C2q Binding Site on Rinusan, a Chimeric Antibody with a Human IgG1 Fc. J Immunol 2000; 164:4178-84.
94. Richardson JS, Richardson DC. Natural b-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 2002; 99:2754-9.