Models for antibody behaviour in hydrophobic interaction chromatography, and in self-association

DOI:
10.1016/j.xphs.2018.11.035

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Journal of Pharmaceutical Sciences

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher’s definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Models for antibody behaviour in hydrophobic interaction chromatography, and in self-association.

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Journal of Pharmaceutical Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>18-848.R1</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Research Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>01-Nov-2018</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Hebditch, Max; University of Manchester, School of Chemistry Roche, Aisling; University of Manchester, School of Chemical Engineering and Analytical Science Curtis, Robin; University of Manchester, School of Chemical Engineering and Analytical Science Warwicker, Jim; University of Manchester, School of Chemistry</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Biotechnology, IgG antibody(s), Biophysical model(s), Protein aggregation, Molecular modeling, Physicochemical properties, Protein(s), Self-association, Computational biology, Interaction(s)</td>
</tr>
</tbody>
</table>
Title

Models for antibody behaviour in hydrophobic interaction chromatography, and in self-association.

Authors

Max Hebditch 1,#, Aisling Roche 2,#, Robin A. Curtis 2, Jim Warwicker 1,*

1 School of Chemistry, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

2 School of Chemical Engineering and Analytical Science, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

# These authors contributed equally to this work

* Correspondence to Jim Warwicker (E-mail: jim.warwicker@manchester.ac.uk; Telephone: +44-161-306-4490)
Abstract

Monoclonal antibodies (mAbs) form an increasingly important sector of the pharmaceutical market, and their behaviour in production, processing, and formulation is a key factor in development. With datasets of solution properties for mAbs becoming available, and with amino acid sequences, and structures for many Fabs, it is timely to examine what features correlate with measured data. Here, previously published data for hydrophobic interaction chromatography (HIC) and the formation of high molecular weight species (HMWS) are studied. Unsurprisingly, aromatic sidechain content of complementarity determining regions (CDRs), underpins much of the variability in HIC data. However, this is not reflected in non-polar solvent accessible surface enrichment at the antigen combining site, consistent with a view in which hydrophobic interaction strength is dependent on curvature as well as extent of an interface. Sequence properties are also superior to surface-based structural properties in correlations to the HMWS data. Combined length of CDRs is the most important factor, which could be an indication of flexibility that facilitates CDR-CDR interactions in mAb self-association. These observations couple to our understanding of protein physicochemical properties, laying the groundwork for improved developability models.

Keywords: Biotechnology; IgG antibody; Biophysical model; Protein aggregation; Molecular modelling; Physicochemical properties; Proteins; Self-association; Computational biology; Interactions

Abbreviations used: HIC, hydrophobic interaction chromatography; HMWS, high molecular weight species; LMWS, low molecular weight species; SASA, solvent accessible surface area; CDR complementarity determining region; FDA, Food and Drug Administration; EMA, European Medicines Agency; PDB, Protein Data Bank; NIST, National Institute of Standards and Technology; KEGG, Kyoto encyclopedia of genes and genomes; NPP, non-polar to polar; SMAC, standup monolayer chromatography.
Introduction

Protein stability in solution can be a major issue for the development of biotherapeutics. Both the inherent protein fold stability and self-association between protein molecules contribute to physical instability. Whether reversible or irreversible, self-association can lead to undesirable solution behaviour such as phase separation or aggregation which can limit, or even abrogate, the development of a promising biopharmaceutical lead. Monoclonal antibodies (mAbs) are typical biotherapeutics. Their large size in comparison with the target binding region they bear, giving a low potency, commonly contributes to a requirement that high protein concentration is required to deliver an effective dose. This high concentration requirement exacerbates issues with antibody self-association, the overcoming of which is therefore a target in the development of therapeutic mAbs.

Many physicochemical factors are implicated in protein self-association, but a key contribution often derives from the presence of hydrophobic moieties on the protein surface, which has been studied both experimentally and theoretically, and is also the basis for predictions. Electrostatic interactions can also play a role, as can a combination of electrostatics and hydrophobicity. Previous work from our group has focussed on the identification of hydrophobic and charged surface patches using electrostatic calculations. Other theoretical approaches to improving antibody solution behaviour have invoked the role of hydrophobicity in the self-association of antibodies. However, theoretical approaches for understanding and predicting antibody self-association have been limited by a lack of publically available datasets. Since commercial potential is embedded in the solution behaviour of biotherapeutics, such data is generally protected through the development process. Significant work has been undertaken where access to large datasets is possible, for example the development by Lonza of a tool for predicting aggregation potential. However, now that the field is mature, with more than 50 mAbs in the clinic, this situation is changing, since the research community has access to the mAbs, their amino acid sequences, and in many cases also the fragment antigen binding (Fab) structures, often in complex with an antigen (drug target).

In this study, a dataset of 24 mAbs is examined, from experimental data reported for 30 protein therapeutics (mAbs, Fc-fusion proteins, antibody drug conjugates, and one bispecific antibody). Focussing on mAbs allows bioinformatics analysis of a single set of complementarity determining regions (CDRs) for each therapeutic. This dataset of commercialised biotherapeutics includes reported retention times in hydrophobic interaction chromatography (HIC), along with the percentage of high molecular weight species (HMWS). Reports indicate that mAbs with greater retention in HIC studies may also exhibit greater precipitation, but also that such correlation is weak. Other studies of HIC data in the
context of sequence and structure indicate that low hydrophobicity correlates with reduced retention time. Recent work attempts to predict solvent accessible surface area (SASA), and associated non-polar area, from amino acid sequence, and finds reasonable performance of a binary classifier for lower and higher retention times in HIC.

In the current work, we use our recently reported sequence-based and structure-based methods for analysing properties that correlate with the experimental data. We have previously shown that whilst CDRs tend to be rich in amino acids with sidechains containing aromatic groups, this non-polarity is not evident when viewed with a standard measure of SASA. It was inferred that this apparent conflict underpins important and specific properties of mAbs that may contribute to not only their high binding affinity, but also to the range of solution and developability properties observed. With the available HIC data, we again find that whilst the aromatic content (a sequence-based property) correlates with experimental variation, structural properties are less evident. For HMWS data, a correlation with combined CDRs length suggests that protein flexibility may play a role. Although significant correlations are found, upon which to base prediction schemes, we suggest that there exists scope for improvement of developability models through consideration of the geometry and flexibility of mAb antigen combining sites.

**Materials and Methods**

*Experimental data, mAb amino acid sequences and Fab-antigen complex structures*

Measurements were obtained from the literature. Data includes HIC apparent retention factors for 24 mAbs, using an alkyl amide column, and percentage of HMWS for 21 mAbs. All of these mAbs are Food and Drug Administration (FDA) and European Medicines Agency (EMA) approved therapeutics. Data on low molecular weight species (LMWS) were not analysed in the current study, since it is directed at understanding the molecular determinants of interaction, with hydrophobic media and via self-association, rather than fragmentation.

Amino acid sequences for heavy and light chains of mAbs (or Fabs) were obtained from various sources: the Drug Bank, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, ChEMBL, the Protein Data Bank, PDB, from the National Institute of Standards and Technology (NIST) for the NIST reference mAb (https://www-s.nist.gov/srmors/view_cert.cfm?srn=8671), and from patent records for Reslizumab (US patent number 5998586). For 6 of the 24 mAbs, the sequences were obtained from Fab data. In these cases (Ofatumumab, Infliximab, Belimumab, Palivizumab, Eculizumab,
Panitumumab), the full light chain sequences were paired with the Fab heavy chain sequences. Amino acid sequences, and their source databases for all 24 mAbs (or Fabs) are given in Supplementary file 1.

For sequence-based calculations, heavy (or VH and CH1 for Fabs) and light chain sequences were combined for each mAb. CDR sequences were extracted from heavy and light chains using a set of sequence-based rules (http://www.bioinf.org.uk/abs/) \textsuperscript{32}, and concatenated to give a single entry for each mAb. In looking for structural representatives, only coordinate files that include a bound antigen are included, so that antigen binding regions on the Fab can be compared with other surface regions. For 3D structural analysis, identifying Fab atoms that are in contact with an antigen focusses more closely on the binding site than does simply including all atoms within the CDRs. A search of the PDB found Fab-antigen complexes for 16 of the 24 mAbs, which are denoted along with the corresponding PDB coordinate files in Supplementary file 1. This supplementary data file also contains the concatenated CDR sequences, which vary in length from 58 to 70 amino acids (for one copy of VH and VL domains).

**Sequence-based calculations**

Thirty-five sequence-based properties were calculated as deviations from population average values for the heavy and light chain combinations, and for the concatenated CDRs. These properties are the 20 amino acid percentage compositions and 15 derivative features (e.g. predicted net charge) that are used in the protein-sol server for prediction of protein solubility (www.protein-sol.manchester.ac.uk). Briefly, the derivative features and their short names are: 7 composite compositions (with single letter amino acid codes representing percentage composition): K-R (KmR), D-E (DmE), K+R (KpR), D+E (DpE), K+R-D-E (PmN), K+R+D+E (PpN), F+W+Y (aro); and 8 further predicted features: pI, hydropathy (mem), charge predicted at pH 7 (chr), fold propensity (fld), predicted disorder (dis), sequence entropy (ent), β-strand propensity (bet), length (len) \textsuperscript{17}. Population averages refer to the \textit{E. coli} dataset \textsuperscript{33} that was used for training the protein-sol server. Here, they are used simply as reference values for deviations, and the ranking of values between different mAbs will be the same for absolute values and deviations. Calculations for the 24 mAbs were made with an offline version of the protein-sol code, available for download at www.protein-sol.manchester.ac.uk, suitable for multiple sequence input.
Structure-based calculations

Coordinate files were each inspected and edited to provide a single copy of a Fab and an antigen bound at its CDRs. Following previous work \(^{18}\), surface and Fab-antigen interface regions were defined according to SASA, and SASA change upon antigen binding. In order to investigate how surface polarity may relate to the HIC and HMWS data, a ratio of non-polar to polar (NPP) SASA is obtained for patches centred on each non-hydrogen atom of a Fab. Each patch includes contributions to SASA from all other non-hydrogen atoms within 13 Å of the central atom. This patch radius typically produces around 1000 Å\(^2\) of SASA, a value representative of the SASA buried at an interface by one of the two protein components \(^{34}\). Similar patch analysis has been used previously in characterisation of interfaces \(^{35,36}\). Two properties are recorded for a Fab, an average of the NPP ratio for all atoms within a classification of surface or antibody-antigen interface, and the maximal NPP ratio within the same classifications. Following previous work \(^{18}\), Fab atoms at the heavy to light chain interface are excluded from the surface classification. Protein surfaces, colour-coded by NPP ratio, are visualised in the ‘patches’ tool of the protein-sol server \(^{18}\), using the NGL graphics software \(^{37}\). Custom visualisation, including for cleft propensity, uses an offline version of the protein-sol software, and PyMOL (Schrödinger L.L.C.).

Data analysis

Correlations between experimental measures (HIC apparent retention factor and percentage of HMWS \(^{22}\)), and sequence-based or structure-based properties were calculated and visualised with heat maps, using Excel and Python. Statistical significance of the Pearson correlation coefficient between a calculated feature and a measured property is assessed at \(p=0.05\). Where (N) multiple tests are made, a Bonferroni correction is applied, such that a \(p\) value of \(< 0.05/N\) is required for significance. The Bonferroni correction compensates for the testing of multiple hypotheses, which here is correlation between experimental data and one or more of the multiple sequence-dependent features.

Results

Aromatic sidechain content of CDRs correlates with retention in HIC

For 24 mAbs, HIC retention and HMWS data were compared with deviations from population averages for 20 amino acid compositions and 15 further sequence-derived properties used
The deviations are signed, so that any correlations observed reflect the underlying trend of, for example amino, acid composition. Calculations were made for concatenations of the CDRs within a Fab, and for combined heavy and light chain sequences (complete where possible, but including only VH and CH1 domains for 6 of the 24 heavy chains, see Materials and Methods). Figure 1 shows a heatmap of HIC and HMWS data correlations with the 35 sequence-based properties, for concatenated CDRs and combined heavy and light chain sequences. Requiring statistical significance at p<0.05 for a correlation of 24 data pairs, Bonferroni corrected for the 35 tests being made, gives a threshold correlation coefficient (r) of 0.62. Of all the comparisons shown in Figure 1, only the correlation of aromatic sidechain content for CDRs and HIC data exceeds this threshold (r=0.639), with a scatter plot shown in Figure 2a.

There are no sequence-based features that individually give statistically significant correlations to HMWS data. Isoleucine content gives a consistent negative correlation in Figure 1, but the lack of statistical significance is shown here with a drop in the magnitude of r from -0.44 to -0.09 (isoleucine content versus HMWS data) when the data point with highest percentage of HMWS is removed (Pembrolizumab, 8.8%). Correlation of valine composition in CDRs with HMWS data behaves in a similar manner to that of isoleucine upon removal of the single (Pembrolizumab) data point, decreasing in magnitude (r=-0.45 to r=-0.16). By comparison, correlation between HMWS data and concatenated CDRs length only falls from 0.44 to 0.41 when Pembrolizumab is removed. The CDR-H3 loop is generally the most variable in length, but here has a correlation coefficient of only 0.19 with the HMWS data.

Solubility predictions from the protein-sol server give correlation coefficients of the expected sense (negative correlation), with HIC and HMWS data, of -0.35 and -0.26, respectively, but are not significant at the 5% level, even with the Bonferroni correction removed for this single test aggregated over 10 of the 35 sequence-based features. It is encouraging though that both aromatic sidechain content and protein length are in the 10 test feature subset for protein-sol. However, the lengths of concatenated CDRs are systematically shorter than the population average protein length used for protein-sol training, illustrating that solubility prediction based on loops (CDRs) is not expected to align precisely with that based on whole protein.

**Non-polar solvent accessible surface area at the antigen interface does not correlate with HIC or HMWS data**

For structure-based calculations, just two properties are being tested against the experimental data, the NPP ratio maxima at firstly the antigen binding interface, and
secondly on the Fab (non-interfacial) surface, and a Bonferroni correction is not applied. Then, significance at the 5% level is achieved when correlation coefficients are greater than 0.50 (for the 16 Fab structures matched to HIC data) and 0.54 (14 Fab structures matched to HMWS data). The correlation between aromatic sidechain content of CDRs and retention in HIC is not replicated with correlation of HIC data to interfacial non-polar area. Specifically, the maximum of NPP ratio, on a Fab at the antigen binding interface, has a correlation coefficient of just 0.09 with the HIC data (and 0.13 with HMWS data). In contrast, surface (non-interfacial) NPP ratio maxima gives a correlation coefficient of -0.51 with HMWS data (significant), and -0.09 with HIC data.

Combining one sequence-based and one structure-based property yields correlation with HMWS data

From the sequence-based analysis, CDRs combined length is the most promising feature for investigating the variation in HMWS data, and for structure-based calculations the most promising feature in this regard is the surface (non-interfacial) maximum NPP ratio. A linear combination of these two features gives \( r = 0.59 \) (\( p = 0.026 \) for 14 data points, Figure 2b), for correlation with HMWS data. The linear combination uses the minima and ranges of the two calculated properties, over the 14 data points, to scale each onto a range of 0 to 1. The sign of the scaled surface maximum NPP ratio is then inverted to convert a negative correlation with HMWS data into a positive correlation, and this value added to the scaled value for CDRs combined length. In line with the improvement shown upon combination, a comparison between maximum surface NPP ratio and CDRs length gives a correlation coefficient of just 0.25. In addition to those two features, 3 other calculated features individually gave correlations with HMWS data of \( r > 0.4 \), the contents of isoleucine, valine, and glutamine. However, when any one of these features was scaled and added to the linear combination of Figure 2b, the overall correlation to HMWS data decreased in each case.

Discussion

Molecular models for the retention of mAbs in HIC and for self-association in forming HMWS

In this study of sequence and structural features that correlate with experimental measures of mAb behaviour, we have found a single dominant feature that correlates with HIC data (Figure 2a), and a combination of two features that gives a correlation with HMWS data
(Figure 2b). The work is enabled by the availability of medium-throughput measurements of characteristics associated with mAb developability. These data have been collected from commercially available samples of the mAbs, and therefore each mAb solution will be subject to a potentially different expression and purification protocol. We have been able to assess the impact of such variation for HIC data, since an equivalent measurement is made in a further report describing a range of biophysical measurements on clinically relevant mAbs. In this latter study, the VH and VL regions of mAbs are reformatted for HEK cell expression in a uniform IgG1 isotype, and with a standardised purification, thereby reducing potential issues of heterogeneity. We calculate the Pearson correlation coefficient between HIC measurements for the two studies at 0.92, showing that, for HIC at least, variation from expression and cell line differences is small. For HMWS, there is no directly equivalent measurement in the study of reformatted mAbs, but the feature that correlates best in the standardised HEK expression study is standup monolayer chromatography (SMAC), with r = 0.66. SMAC is reported to be effective in assessing the developability of protein therapeutics.

Although elements from general prediction schemes, such as the aromatic sidechain content calculated with the protein-sol server, are relevant in correlations with experimental HIC and HMWS data, for mAbs the CDRs play a particularly important role. The current report extends previous computational work, facilitating more detailed discussion of what is the physicochemical basis for correlation. For example, the lack of correlation between extent of non-polar SASA at the antigen binding site of a mAb and HIC or HMWS data, promotes discussion of whether a simple measure of SASA gives a correct assessment of non-polar interactions.

Considering retention in hydrophobic interaction chromatography, it is not surprising that the content of amino acids containing aromatic sidechains is positively correlated, in line with the relative hydrophobicity of tyrosine, phenylalanine and tryptophan. Indeed, these amino acids are enriched at protein-protein interfaces, and also at the antigen binding sites of mAbs. So, why does the composition of these amino acids not also correlate well with HMWS data? Accessibility to the aromatic binding surfaces will play a role. An antigen binds with steric complementarity to a mAb (Figure 3a), and flexible alkyl chains on the support in HIC are also likely to be able to access the antigen binding site on a mAb (Figure 3b). In this model, another mAb, with ordered CDRs adapted for binding antigen, will be a mismatch for a copy of its own binding surface, and will therefore not be so effective in accessing the aromatic binding surfaces (Figure 3c). Thus correlation with HIC data may not lead to a correlation with HMWS data, due to differences in the flexibility of molecular species scanning a binding site formed by the CDRs.
Having proposed that rigidity within mAb CDRs could deflect scanning of hydrophobic interactions by a neighbouring mAb, a positive correlation between length of combined CDRs and HMWS data suggests that nevertheless CDRs are, on average, playing a role in self-association, consistent with observation that Fab-Fab interaction mediates self-association 42. Since longer loops in mAb CDRs are inherently more flexible 43, it could be that increased CDRs length is linked to more flexible structure with the potential for partial unfolding and disordering, and interaction with the similarly flexible CDRs of a neighbouring mAb (Figure 3d). Antibody bivalency would then facilitate formation of HMWS. Partial unfolding of proteins is generally considered as a risk factor for formation of oligomers via non-specific interactions, including in antibody solutions 44. This is not to say that the CDRs are predominantly unfolded, rather that partial unfolding in a sub-population would lead to self-association.

The second feature (in addition to CDRs length) used in the linear fit of Figure 2b is the maximum surface (non-interfacial) NPP ratio of a Fab. The maximum NPP ratio determined from a static structure at the Fab-antigen interface will play less of a role in a scheme for mAb-mAb interaction with flexible, partially unfolded CDRs. If interactions between flexible CDRs lead to HMWS, then this interaction could be enhanced through biasing of configurations towards CDR-CDR contacts. Such biasing could be achieved in part by reducing maximal surface NPP ratio away from the CDRs, consistent with the observed negative correlation to HMWS data.

Properties of mAbs that have not been included in the current calculations are known to influence developability, for example glycosylation and its heterogeneity 45, and chemical modification of sidechains 46. The central role of CDRs in providing correlations in this study may be reduced as other biophysical measurements are considered. Whilst predictive schemes for oxidation, based on SASA, are promising 46, the flexible nature of glycosyl groups presents a major challenge for predictive modelling.

Assessment of non-polar interfaces and hydrophobic interaction

In previous work we noticed that whilst the interface between the heavy and light chains of Fabs conformed to the view that non-polar patches are generally important in protein-protein interactions 39, antibody-antigen surfaces are relatively polar 18. Assessment of non-polar or polar is made with a patch-based measure of surface at the binding site that is exposed when the partner protein is removed. Our previous result is recapitulated here with an apparent conflict between enrichment in CDRs for amino acids with aromatic sidechains (which correlates with HIC data), but no correlation evident when non-polar surface of the interface to antigen is calculated. That many other interfaces, such as that between heavy
and light chains, are relatively non-polar can be viewed for example with the ‘patches’ option of the protein-sol server (www.protein-sol.manchester.ac.uk). Thermodynamic analysis of mutations shows that the hydrophobic effect is important for mAb-antigen interactions. It can therefore be concluded that simple geometrical patch-based estimation of non-polar SASA to polar SASA ratio is an insufficient measure for assessing hydrophobicity of mAb-antigen interactions. The antigen binding surface of an exemplar Fab from the dataset (Ipilimumab Fab, PDB id 5tru) is shown in Figure 4a, and the more non-polar patches of the heavy and light chain interfaces for the same Fab in Figure 4b.

Within the content of aromatic sidechains, tyrosines are numerous. Typically the hydroxyl group is more solvent exposed than the aromatic ring for tyrosine sidechains. To estimate the role of this differential accessibility in calculated NPP ratio, the surface of Figure 4a was recalculated, but with all tyrosines mutated to phenylalanines (lacking the hydroxyl group, Figure 4c). The antigen binding surface of the Fab becomes significantly more non-polar. The maximum NPP ratio for patches in the Fab binding site for antigen, increases from 1.85 in Figure 4a to 3.70 in Figure 4c. Corresponding SASA values for these patches for wild type are 654.5 Å² non-polar, 353.6 Å² polar (Figure 4a), and 848.6 Å² non-polar, 229.2 Å² polar for mutated (Figure 4b). This alchemical calculation gives the insight that more exposed hydroxyl oxygen atoms contribute more to SASA than less exposed aromatic carbon atoms. Although this is easily understood in terms of the relative degrees of solvent exposure of these atom types, the consequences for SASA-based calculations are important. Estimation of area-based hydrophobic binding energy depends on the number of water molecules adjacent to the molecular surface, in a solvent accessible surface that is drawn out by the water probe centre. Thus, the solvent accessible surface is located beyond the molecular surface. At more solvent accessible sites, for example tyrosine hydroxyls, SASA will be larger than the corresponding molecular surface area. For less solvent accessible sites, typical for aromatic groups in sidechains, the inverse is true. Thus what might seem like a relatively non-polar surface in a view of molecular surface colour-coded by atom type, may not translate to a relatively non-polar solvent accessible surface (or NPP ratio). Since the solvent accessible surface is the core of standard analyses of the hydrophobic effect, there remains the question of how such models handle the antigen binding sites of antibodies.

Mutation analysis suggests that strength of the non-polar SASA-dependent hydrophobic effect varies across an antibody-antigen interface. Phenylalanine at a central part of an antibody binding site (with lysozyme) was replaced with smaller hydrophobic amino acids. From the resulting thermodynamic analysis the authors estimate that hydrophobic interaction for an empirical SASA model is 46 cal mol⁻¹ Å⁻² at the centre of the binding site, compared with 21 cal mol⁻¹ Å⁻² at the periphery. The authors of these
mutation studies suggest that the hydrophobic effect is dependent on environment and curvature, consistent with earlier suggestions from theoretical analysis 50.

Our results, i.e. lack of correlation between structure-based NPP ratio and either HIC data or aromatic content of CDRs, are consistent with an environment-dependence of the hydrophobic effect. This is not surprising since a modelled proportionality between non-polar SASA buried in an interface, and a component of the binding energy, is an empirical construct to approximate hydration effects in the binding of non-polar surfaces. However, this empirical model generally performs well in identifying protein-protein interfaces 18, so that its failure for antibody-antigen interfaces is noteworthy.

Part of the answer may lie in the observation that CDR regions generally have a geometry that is somewhat less flat than a typical protein surface, but not as cleft-like as an enzyme active site 51. In the context of an environment-dependent hydrophobic effect 47,50, the more cleft-like antibody-antigen interface would effectively have a higher scaling of non-polar SASA to binding contribution. We have previously developed a computational measure of cleft-like geometry in a protein 51,52, which is shown here in a view of the antigen binding surface (Figure 4d). An increased contribution to binding from the regions with lower solvent accessibility highlighted in Figure 4d could be incorporated, if a suitable scale can be devised. Derivation of a scale would include binding data 47,49, and theoretical methods for entropy calculation in protein-solvent systems 53,54.

Antibody-antigen interfaces are atypical protein-protein interfaces, since they do not develop through co-evolution of the partner sites. Antigen surfaces exhibit a range of polarities and do not evolve non-polar surface to accommodate antibody binding. It is possible that antibodies have adapted to mitigate this issue through the use of binding surfaces intermediate between flat and the degree of cleft found in enzymes, consistent with observed differences between antibody-antigen interfaces and other protein-protein interactions 55. This may effectively allow for rescue of a larger non-polar contribution to binding energy, where a lack of high NPP ratio is compensated by enrichment for binding pockets and enhanced, environment-dependent, hydrophobicity, possibly underpinning the discovery of binding hot-spots at antibody-antigen interfaces 56.

Conclusions

If our understanding of the energetics of antibody-antigen binding is still developing, what are the implications for developing predictive models of antibody solution behaviour? Models have tended to focus on non-polar SASA and charge. Net charge repulsion is known to modulate colloidal stability, additionally favourable interaction of asymmetric
charge distributions may also be important. For this dataset, we find that predicted net charge of whole mAbs or the combined CDRs does not give statistically significant correlation with HIC or HMWS data. However, since a large majority of the antibodies have similar basic isoelectric points, there is probably insufficient dynamic range to identify a relationship. Other work has found that the charge carried by CDRs contributes to an understanding of association measured with self-interaction nanoparticle spectroscopy.

The elemental models for antibody behaviours in HIC and in self-association (formation of HMWS) schematised in Figure 3, and based on the correlations in Figure 2, will be refined when more data is incorporated. At this stage, the models clearly suggest aspects for further study. With regard to hydrophobic interaction measured in HIC, aromatic content of CDRs sequence is important, but also raising the challenging question of how to read this in 3D. A related issue is how to assess the strength of curvature-dependent hydrophobic interactions. For self-association, two observations are made, but need more wide-scale analysis to confirm their generality. Does polarity of antibody surface outside of the CDRs contribute to determining self-association through biasing towards CDR–CDR mediated configurations? Within the CDRs, what contribution does length (and potential flexibility) make to the ability to form CDR-CDR mediated interactions? As these questions are addressed, it should be possible to deliver more accurate predictive models for antibody developability.

Acknowledgements

The authors would like to acknowledge funding from the UK EPSRC (grant EP/N024796/1), and from the EU Horizon 2020 Research and Innovation programme (Marie Sklodowska-Curie agreement 675074).
References


Figure Legends

Figure 1. A heat map of correlations between sequence properties and measurements. Pearson correlation coefficients are colour-coded, for comparison with either the HIC or HMWS data. The 35 sequence-based properties are: 20 amino acid compositions (given by single letter code), 7 composite compositions K-R (KmR), D-E (DmE), K+R (KpR), D+E (DpE), K+R-D-E (PmN), K+R+D+E (PpN), F+W+Y (aro); and 8 further predicted features: pI, hydropathy (mem), charge predicted at pH 7 (chr), fold propensity (fld), predicted disorder (dis), sequence entropy (ent), β-strand propensity (bet), length (len) 17. Two sets of correlations are shown, either for the concatenated sequences from the combined CDRs of a Fab (CDR), or for concatenation of the light and heavy chains (LH), in some cases lacking the CH2 and CH3 domains.

Figure 2. Features that correlate with HIC data (retention times in the chromatography) and, separately, HMWS data (percentage of high molecular weight species). (a) Content of amino acids containing aromatic sidechains (FWY), in the combined CDR sequences, is sufficient to provide a significant correlation with HIC data. (b) For HMWS data, a significant correlation is seen for a linear combination of two features, one sequence-based (CDRs length), and one structure-based, maximum surface (non-interfacial) NPP ratio. In both panels (a) and (b) the line of best fit and correlation coefficient are shown, and symbols are colour-coded to allow identification of mAbs from the list that is included in the Figure.

Figure 3. Schematic models for mAb behaviour in HIC and in formation of HMWS. (a) A cartoon view of antibody binding to antigen is shown, emphasising surface complementarity. (b) Flexible hydrophobic groups of the HIC column would be able to access the binding pocket, probing hydrophobicity in the pocket and the composition of aromatic groups. (c) With a second antibody (CDRs) replacing the antigen, access to binding pockets is restricted due to lack of complementarity for the rigid proteins. (d) A population of partially unfolded CDR loops could facilitate interaction between antibodies, but the specific binding pockets would be lost, and dependence on aromatic composition reduced. Instead, such interaction (and potentially the formation of HMWS) would be higher for longer and more flexible CDRs. Interaction strengths are denoted by ✔✔ (strong), ✔ (weaker), X (weakest).

Figure 4. The non-polar surface and cleft geometry of CDRs. (a) A view into, and centred on, the CDRs shows no great non-polarity (green is relatively non-polar, purple relatively polar, white is intermediate), for the exemplar Fab of Ipilimumab, PDB id 5tru. (b) Views into surfaces that form the heavy-light chain interface show relatively non-polar patches. (c) With
tyrosine residues mutated to phenylalanine, the CDRs-centred view is now relatively non-polar. (d) Colour-coding according to a cleft index 52, with blue/cold most exposed and red/hot least exposed, highlights binding pockets formed by the CDRs.

Supporting Information Available

Supplementary file 1: An Excel format file contains information for each of the 24 mAbs studied: mAb name; whether the full heavy chain sequence or only the Fab heavy chain sequence was retrieved; source for the heavy and light chain sequences, reference number in that source, and URL; Fab-antigen complex PDB id (if available, N/A otherwise); heavy chain sequence, light chain sequence, concatenated CDRs sequence.
Figure 1

1058x264mm (72 x 72 DPI)
Figure 2

1461x600mm (72 x 72 DPI)
Figure 3

a) mAb-Ag

b) mAb in HIC

c) mAb-mAb

d) mAb-mAb, partially disordered CDRs
null
NSSHIDYADSVQGVSLTASSLDY
SGSTYADDVSKGRHPGPGFVDY
YTSFETYADDKREPSKPYQGGSHWYFDV
GINTDNFPEFTALSYYDFAFY
DSSSTNYAPSLKDPDGNNYWFVDY
NSGSGYADDVSKGDIQSYNYGFDGY
SGGSGYNQFQKLNSPFSYFDY
SSYYADSVKSYDFAFV
GDSYNQKRGSTYSGGDFWYFPV
GITYADDVSKWGGDFHYAMDY
SGGSTYADDVKGDIPGTTVMSWDF
NPMYSITDDYQRFKGTYDFGTGVV
GSSHYADSVKGDNYDGY
VSQGGGTNRENEFMRQYRFDMGFV
DGNNKYYADSVKSTGWGPPFDY
GITYQDPRQSEGTYSHGQAMDY
KSNANTRYAEVGNYGTDFV
IPMFTGATYQSNQGSSRDLILLPHHALSP
IFPGDDTDDYNSQFRGVSDFGTVLYV
WDQDETYNSPUSKSNMTAGYFDV
WDDKDYNSLXDDMIFNYFDV
GSGADVETKDFFFSPSWYFDV
YSGNTYNSLSDVDTGAFDQ
QTDYNSAIKSEYYYGYFDV