Data mining patented antibody sequences

Konrad Krawczyk, Andrew Buchanan, and Paolo Marcatili

Abstract

The patent literature should reflect the past 30 years of engineering efforts directed toward developing monoclonal antibody therapeutics. Such information is potentially valuable for rational antibody design. Patents, however, are designed not to convey scientific knowledge, but to provide legal protection. It is not obvious whether antibody information from patent documents, such as antibody sequences, is useful in conveying engineering know-how, rather than as a legal reference only. To assess the utility of patent data for therapeutic antibody engineering, we quantified the amount of antibody sequences in patents destined for medicinal purposes and how well they reflect the primary sequences of therapeutic antibodies in clinical use. We identified 16,526 patent families covering major jurisdictions (e.g., US Patent and Trademark Office (USPTO) and World Intellectual Property Organization) that contained antibody sequences. These families held 245,109 unique antibody chains (135,397 heavy chains and 109,712 light chains) that we compiled in our Patented Antibody Database (PAD, http://naturalantibody.com/pad). We find that antibodies make up a non-trivial proportion of all patent amino acid sequence depositions (e.g., 11% of USPTO Full Text database). Our analysis of the 16,526 families demonstrates that the volume of patent documents with antibody sequences is growing, with the majority of documents classified as containing antibodies for medicinal purposes. We further studied the 245,109 antibody chains from patent literature to reveal that they very well reflect the primary sequences of antibody therapeutics in clinical use. This suggests that the patent literature could serve as a reference for previous engineering efforts to improve rational antibody design.

Introduction

The binding versatility of antibodies has made them the most successful group of biotherapeutics. Typical timelines involved in bringing these molecules to the market are long, but more and more molecules are approved in the US and EU each year. Successful exploitation of antibodies by either experimental or computational techniques relies on our ability to understand what makes a successful antibody-based therapeutic.

Therapeutic antibodies on the market and in late-stage clinical trials have been previously studied by experimental and computational approaches to identify properties that make a successful biotherapeutic. One such study, however, only focused on 137 approved or post-Phase 1 antibodies (Clinical-Stage Therapeutics (CSTs)), which is a small dataset in the light of the mutational space available to antibodies. CSTs are high-quality data-points that are the end results of a long process of selecting a molecule from a number of viable candidates. The single successful therapeutic molecule is therefore only partially representative of the entire engineering process. Full public disclosure of the efforts involved in developing a therapeutic antibody, including intermediate sequences and details regarding selection decisions, is not common because of the commercial value of such know-how, which needs to be legally protected.

To protect the know-how involved in engineering therapeutic antibodies, relevant information needs to be disclosed in patent documents. Previous approaches to extract information about the patent antibody landscape or specific antibody formats have focused on keyword and patent classification searches. One can broadly discern between patents on antibody techniques (e.g., phage display, humanization) and novel antibody molecules. It is the patents on novel molecules that could be of particular engineering interest, as these reflect the constructs that might find their way into the clinic. The disclosure of antibody sequence and target information in such patents reveals to a certain extent the engineering choices because such molecules have been subjected to myriad prior tests to be suitable candidates for expensive legal protection and further clinical trials.

The purpose of the patent literature is not to convey scientific knowledge, but to provide legal protection. In this work, we assessed whether patent data could be useful for therapeutic antibody engineering efforts by establishing the extent to which antibodies described in patents reflect therapeutics in clinical use. For this purpose, we identified patent documents that contained antibody sequences, to quantify how many of these were destined for medicinal purposes and how well they reflect advanced stage therapeutics.
Results

Antibodies account for a non-trivial proportion of sequences deposited in patent documents

We identified documents with antibody sequences by downloading data from four data sources: United States Patent and Trademark Office (USPTO, http://uspto.gov), World Intellectual Property Organization (WIPO, http://wipo.int), DNA Data Bank of Japan (DDBJ), and The European Bioinformatics Institute (EBI). The choice of the data sources was motivated by the availability of biological sequences and coverage of patent documents worldwide. Biological sequence information is not universally available in patent documents in all jurisdictions. In certain cases, the data is not freely available, but rather accessible for a fee (e.g., European Patent Office (EPO)). Primary access to biological sequences in machine-readable format is freely available from the USPTO and WIPO. USA is the largest pharmaceutical market, compelling pharmaceutical companies developing a novel antibody therapeutic to seek patent protection within the jurisdiction of USPTO. Similarly, it is common to seek protection under the auspices of WIPO Patent Cooperation Treaty (PCT) system to spread the coverage of the patent documents across many jurisdictions worldwide. Furthermore, data from certain major jurisdictions, such as EPO, Japanese Patent Office and Korean Patent Office are available via third parties such as DDBJ and EBI.

Therefore, we argue that datasets made available via USPTO, DDBJ, WIPO, and EBI provide reasonable coverage of the worldwide antibody sequence patents.

We extracted raw sequence data from USPTO, WIPO, DDBJ, and EBI on January 30, 2020, with the particulars of parsing the heterogenous sources described in Methods. From each dataset, we extracted raw, redundant amino acid and nucleic acids sequences. Sequences containing exclusively nucleotides were translated to amino acids using IgBlast as described previously. Raw amino acid sequences were analyzed using ANARCI to identify antibody variable region chains (V_H, V_L, including single-chain variable fragments (scFvs)). We report the number of raw sequences analyzed and the resulting identified antibodies in Table 1.

We find a higher proportion of sequences identified as antibodies in amino acid depositions, which account for as many as 10.9% and 12.1% of USPTO-FT and DDBJ datasets, respectively. The majority of sequences deposited in patents are in fact very short. For example, in USPTO-FT only 1,811,694 (32.7%) amino acid sequences are longer than 50 amino acids, and antibodies comprise 30.5% of these. This indicates that antibodies make up a non-trivial volume of all the sequences deposited in patent documents.

Antibody sequence data in patents is, however, redundant to a large extent when one considers a unique sequence to be defined by its variable region. Combining all the non-redundant V_H and V_L sequences from our datasets, we count 245,109 unique antibody domains (135,397 heavy chains and 109,712 light chains). This suggests that many antibody variable region sequences are listed as part of multiple patent documents. Not all of these sequences are guaranteed to have been developed for medical applications, which can be determined by analyzing the text content of patent documents.

Table 1. Published biological sequences and proportion thereof identified as antibody chains. We extracted raw sequences from USPTO (divided between the full text, FT, and long listing repository PSIPS), DDBJ, WIPO and EBI. The total number of raw sequences is given in column Total Raw. Of these we show how many were identified by ANARCI as containing an antibody chain (column Ab-identified). In the column “% Total” we report the proportion of identified antibody sequences out of the total of raw sequences. Both Total Raw and Ab-identified columns report the redundant number of sequences so as to exemplify the volume of antibody depositions in patent sequences – we report the number of unique heavy (H) and light (L) chains in the parentheses in column “Ab-identified”.

| Source      | Sequence Type | Total Raw | Ab-identified (unique Heavy (H), Light (L)) | % Total |
|-------------|---------------|-----------|------------------------------------------|---------|
| USPTO FT    | Amino Acid    | 5,534,127 | (H = 52,388L = 38,922)                    | 10.95   |
|             | Nucleotide    | 7,068,248 | (H = 21,169L = 17,009)                   | 3.24    |
| USPTO PSIPS | Amino Acid    | 25,527,942| (H = 33,806L = 24,086)                  | 1.84    |
|             | Nucleotide    | 176,840,912| (H = 376,567)                        | 0.21    |
| DDBJ        | Amino Acid    | 4,412,209 | (H = 61,999L = 46,015)                  | 12.09   |
|             | Nucleotide    | 44,968,142| (H = 43,185)                           | 0.91    |
| WIPO        | Amino Acid    | 10,275,174| (H = 35,290L = 28,502)                | 4.23    |
|             | Nucleotide    | 13,490,560| (H = 160,542)                         | 1.19    |
| EBI         | Amino Acid    | 10,368,431| (H = 73,450L = 50,326)               | 6.88    |
|             | Nucleotide    | 12,349,772| (H = 38,366)                           | 0.31    |

Patent landscape of documents containing antibody sequences

We analyzed the text content of patents containing antibody sequences to establish what proportion of these list molecules for medicinal purposes. We connected all the redundant antibody sequences to their patent documents and identified a patent family for each. A patent family can be regarded as identifying documents with the same subject matter across several jurisdictions. Altogether our 245,109 sequences are distributed among 16,526 patent families. We extracted the metadata from the patent documents, such as titles, abstracts, inventors, and classifications. We used this information to determine the proportion of patents destined for medicinal applications by analyzing their classifications, and whether the inventors and listed targets resemble entities and molecules associated with development of monoclonal antibody therapies.

Most patent documents citing antibody sequences are destined for medicinal applications

We analyzed the patent classifications of the 16,526 patent families that indicate the purpose of the invention described in each document. We extracted the Cooperative Patent Classification (CPC, developed by USPTO and EPO, https://www.cooperativepatentclassification.org/) designations from the documents, as this was the most common listed scheme, covering 15,951 (96.5%) of 16,526 families. Patent classifications according to CPC
have a section, class, subgroup, main group and a subgroup (e.g., classification C07K16/2866 has section C, class 07, subclass K, main group 16 and subgroup 2866). We divided the 15,951 families according to their CPC classifications excluding the subgroup (e.g., C07K16/2866 becomes C07K16) to reveal the general categories the documents fall into, as shown in Table 2.

Subgroup C07K16, which indicates immunoglobulins, is the most common classification, present in 13,790 (86.5%) of the 15,951 patent families. Families listing antibodies for medicinal purposes (A61K39) account for 9,459 (59.3%) of the 15,951 families. Furthermore, the more general medicinal categorization A61K (preparations for medical, dental, or toilet purposes) accounts for 11,398 (71.5%) of the 15,951 patent families. This indicates that the majority of documents citing antibody sequences are developed for medicinal purposes, such as novel treatments or diagnostics. This is well reflected by the organizations that submit such patent applications, where 9 of the top 10 and 69 of the top 100 are pharmaceutical companies associated with development of monoclonal antibody therapies for a range of targets and disease indications (see Supplementary Section 1).

**Table 2.** Subclasses of the patent classifications. Most common subclasses associated with patents including antibody sequences according to the Cooperative Patent Classification (CPC, https://www.cooperativepatentclassification.org/). There were 15,951 patents containing antibodies with CPC classification and the percentage of families in each class is expressed as a proportion of this number.

| Class | Total families (%) | Description |
|-------|-------------------|-------------|
| C07K16 | 13,790 (86.4) | Immunoglobulins [IGs], e.g. monoclonal or polyclonal antibodies (antibodies with enzymatic activity, e.g. abzymes) |
| C07K2317 | 12,001 (75.2) | Immunoglobulins specific features |
| A61K39 | 9,459 (59.3) | Medicinal preparations containing antibodies or antigens |
| C07K2319 | 3,451 (21.6) | Fusion polypeptide |
| G01N33 | 3,105 (19.4) | Investigating or analyzing materials by specific methods not covered by groups G01N1/00 – G01N3/00 |
| C07K14 | 3,037 (19.0) | Peptides having more than 20 amino acids; Gastrins; Somatostatin; Melanotropins; Derivatives thereof |
| A61K47 | 2,392 (14.9) | Medicinal preparations characterized by the non-active ingredients used, e.g. carriers or inert additives; Targeting or modifying agents chemically bound to the active ingredient |
| A61K38 | 2,058 (12.9) | Medicinal preparations containing peptides |
| C12N15 | 1,972 (12.3) | Mutation or genetic engineering; DNA or RNA concerning genetic engineering, vectors, e.g. plasmids, or their isolation, preparation or purification; Use of hosts therefor |
| A61P35 | 1,900 (11.9) | Specific therapeutic activity of chemical compounds or medicinal preparations |
| A61K45 | 1,671 (10.4) | Medicinal preparations containing active ingredients not provided for in groups |
| A61K31 | 1,415 (8.8) | Medicinal preparations containing organic active ingredients |
| G01N2333 | 1,329 (8.3) | Assays involving biological materials from specific organisms or of a specific nature |
| C12N5 | 816 (5.1) | Undifferentiated human, animal or plant cells, e.g. cell lines; Tissues; Cultivation or maintenance thereof; Culture media therefor |

**Targets of antibodies in patent documents correspond to known therapeutic targets**

We assessed the extent to which antibody targets reported in patent literature reflect those of known therapeutic antibodies. Each patent family in PAD was scanned for antibody target (see Methods). Therapeutic antibodies in clinical use together with their associated targets were compiled from the WHO lists of International Nonproprietary Names (INNs, e.g., list 122(20), the international ImMunoGeneTics information system* (IMGT, http://www.imgt.org), 21 The Antibody Society (http://www.antibodysoociety.org), and Thera-SAbDab,22,23 resulting in 563 unique INNs. We grouped the targets by number of patent families and therapeutic antibodies they were associated with. The results for the top 30 targets sorted by the highest number of patent families are shown in Table 3.

The number of patent families associated with a target appears to correspond to a larger number of therapeutic antibodies against the same target. The top 10 targets sorted by number of their patent families account for 114 (20.2%), the top 30 account for 223 (39.6%), and the top 100 account for 369 (65.5%) of 563 therapeutics. Therefore, targets from patents listing antibody sequences provide a reasonable reflection of the targets of currently available therapeutic antibodies. In fact, the greater number of patent families per target can be associated with an earlier date of the said target being mentioned in a patent document (Figure 1a). It does not mean, however, that the patent space for monoclonal antibodies is saturated, as the number of new targets mentioned is increasing (Figure 1b).

**Table 3.** Top 30 targets in patent documents. We extracted the targets of the antibodies in patent documents and present top 30 ranked by the number of families where they were mentioned. For each target, we show the number of patent families mentioning the target (#Families), the number of therapeutics on the market/in the clinic against it (#Therapeutics) and the cumulative number of therapeutics covered by the top targets (#Therapeutics cumulative).

| Rank | Target | #Families | #Therapeutics | #Therapeutics (Cumulative) |
|------|--------|-----------|---------------|---------------------------|
| 1    | pd1    | 284       | 20            | 20                        |
| 2    | cd3    | 221       | 20            | 40                        |
| 3    | her1   | 190       | 17            | 57                        |
| 4    | pd1    | 189       | 12            | 69                        |
| 5    | trna   | 185       | 6             | 75                        |
| 6    | her2   | 175       | 9             | 84                        |
| 7    | cd20   | 169       | 14            | 98                        |
| 8    | influenza | 151   | 5             | 103                       |
| 9    | cmet   | 136       | 4             | 107                       |
| 10   | vegfa  | 135       | 7             | 114                       |
| 11   | amyloid beta | 129 | 8             | 122                       |
| 12   | hiv    | 123       | 2             | 124                       |
| 13   | ili6   | 102       | 7             | 131                       |
| 14   | cd40   | 102       | 9             | 140                       |
| 15   | cd19   | 102       | 6             | 146                       |
| 16   | cta4   | 97        | 6             | 152                       |
| 17   | iil7   | 92        | 8             | 160                       |
| 18   | igfr   | 92        | 6             | 166                       |
| 19   | pcss9  | 89        | 8             | 174                       |
| 20   | her3   | 87        | 6             | 180                       |
| 21   | rsv    | 73        | 5             | 185                       |
| 22   | cd38   | 73        | 5             | 190                       |
| 23   | tau    | 68        | 5             | 195                       |
| 24   | lag3   | 68        | 7             | 202                       |
| 25   | ox40   | 65        | 6             | 208                       |
| 26   | bcma   | 64        | 1             | 209                       |
| 27   | il23   | 63        | 5             | 214                       |
| 28   | cd47   | 63        | 2             | 216                       |
| 29   | ang2   | 62        | 2             | 218                       |
| 30   | vegfr2 | 56        | 5             | 223                       |
This suggests that studying patent documents including antibody sequences could provide an early indication of their targets, and thus activity in the field of therapeutic antibodies.

**The number of patent documents associated with antibody sequences is growing**

We analyzed the timestamps associated with patents to assess whether there is a growing trend in releasing documents with antibody sequences and what proportion thereof is made up of molecules for medicinal indications. Each patent family lists several dates corresponding to the activity associated with the patent. We noted the earliest and most recent dates for each patent family to reflect the original submission dates and the most up-to-date activity, respectively.

We plotted the earliest dates for each patent family in our dataset, which indicates that the number of patent documents containing antibody sequences is steadily rising (Figure 2). The most recent dates associated with the same patent documents (Figure 2) shows a more acute rise since 2016, which indicates strong activity within the earlier submitted patents. Since not all patent families are explicitly destined for medicinal

![Volume of patent documents including antibody sequences](image_url)
applications, we have plotted the corresponding earliest and most recent dates for the 9,459 documents classified as medicinal preparations containing antibodies (Supplementary Figure 1), which recapitulates the increasing number of patent documents being released.

Increasing patent activity in documents listing antibody sequences for medicinal indications is in line with the rising approval rates for antibody-based biologics. Given that the patents are an early sign of approvals to come, it suggests that we can expect more biologics in the clinics in the foreseeable future. Since the majority of such patent documents are indeed listing antibodies for medicinal purposes, the broad characteristics of the molecules listed in patent documents could provide an indication of the engineering choices in their design.

The sequence landscape of patented antibodies

Antibody sequences found in patent documents could reflect the broad decisions taken by engineers shaping the molecules, but not all of these are destined for medicinal applications. We thus analyzed the broad sequence characteristics of antibodies from patent documents to establish to what extent they are a reflection of therapeutic antibodies in clinical use and vice versa. We performed this analysis by looking at all antibodies from all patents (AllPatAb) and just the subset associated with documents classified as containing antibodies for medicinal applications (MedPatAb). Altogether AllPatAb consisted of 135,397 heavy chains and 109,712 light chains, whereas MedPatAb consisted of 93,067 heavy chains (68.7% of all heavy chains) and 67,667 light chains (67.7% of all light chains).

**Most antibody sequences from patents align to human and mouse germ line regions**

We checked the patterns of organism-specific germ line gene usage in antibody sequences originating from patent documents. Since organism reporting is not consistent in patent documents, we aligned the sequences in PAD to Hidden Markov Models created from IMGT germ line sequences for 15 organisms: human, mouse, alpaca, rhesus, rabbit, rat, pig, cow, macaque, zebrafish, trout, salmon, dog, horse and chicken. For each organism and germ line, we noted the total number of patent antibody sequences aligning to a given germ line, as well as the number of families from which they originated.

We show the number of MedPatAb sequences that aligned to one of our 15 organisms in Table 4, with the corresponding distribution for AllPatAb sequences in Supplementary Table 2. The majority of the unique heavy sequences from patents for medicinal indications align to human germ lines (72.8% of unique sequences), followed by mouse (15.4% of unique sequences). The same holds true for light chains with 67.7% of MedPatAb sequences aligning to human and 19.7% to mouse germ lines. Antibodies aligning to either mouse or human germ lines are most frequently found within protein families. Human-aligned heavy and light chains can be identified in 75.7% and 69.8% patent families, respectively. Mouse-aligned heavy and light chains can be found in 52.2% and 53.9% patent families, respectively. This broad proportion is also reflected in all the antibody sequences from patents (AllPatAb), indicating that the medicinal patent classification does not skew the broad trend of the majority of patented sequences aligning to human and mouse germ lines. The alignment to those two organisms reasonably reflects the human

| Table 4. Most common V-region gene species antibodies from patents aligned to. Antibodies from patent documents destined for medicinal indications (MedPatAb) were aligned to 15 IMGT-derived V region germ lines from human, mouse, alpaca, rhesus, rabbit, rat, pig, cow, macaque, zebrafish, trout, salmon, dog, horse and chicken. We noted the number of patent sequences that aligned to the given species germ line (#Unique Sequences) and the number of patent families (#Patent Families) these originated from.

| Per sequence | Per family |
|--------------|------------|
| **HEAVY CHAIN** | | |
| **Organism** | **#Unique Sequences** | **Percentage** | **Organism** | **#Patent Families** | **Percentage** |
| Human | 67754 | 72.80 | Human | 7070 | 75.69 |
| Mouse | 14326 | 15.39 | Mouse | 4874 | 52.18 |
| Alpaca | 7047 | 7.57 | Macaque | 485 | 5.19 |
| Rabbit | 1313 | 1.41 | Horse | 473 | 5.06 |
| Macaque | 1035 | 1.11 | Alpaca | 403 | 4.31 |
| Horse | 799 | 0.85 | Rabbit | 256 | 2.74 |
| Chicken | 417 | 0.44 | Chicken | 46 | 0.49 |
| Dog | 291 | 0.31 | Dog | 28 | 0.29 |
| Rhesus | 43 | 0.04 | Rhesus | 23 | 0.24 |
| Cow | 30 | 0.03 | Cow | 11 | 0.11 |
| Pig | 9 | ~0 | Pig | 9 | 0.09 |
| Rat | 2 | ~0 | Rat | 3 | 0.03 |
| Salmon | 1 | ~0 | Salmon | 2 | 0.02 |

| **LIGHT CHAIN** | | |
| **Organism** | **#Unique Sequences** | **Percentage** | **Organism** | **#Families** | **Percentage** |
| Human | 45828 | 67.72 | Human | 6312 | 69.76 |
| Mouse | 13320 | 19.68 | Mouse | 4880 | 53.93 |
| Rhesus | 5333 | 7.88 | Rhesus | 2238 | 24.73 |
| Rabbit | 1438 | 2.12 | Rat | 361 | 3.98 |
| Rat | 778 | 1.14 | Rabbit | 240 | 2.65 |
| Chicken | 505 | 0.74 | Chicken | 46 | 0.5 |
| Dog | 220 | 0.32 | Cow | 31 | 0.34 |
| Cow | 213 | 0.31 | Dog | 21 | 0.23 |
| Pig | 17 | 0.02 | Pig | 7 | 0.07 |
| Horse | 15 | 0.02 | Horse | 7 | 0.07 |
focus of antibody development and the rodent antibodies that often serve as a basis for humanized therapeutics.  

**Germline V gene usage of antibodies from patent documents corresponds to a large extent with germline V gene usage of therapeutic antibodies**

Given that the majority of antibodies from patents align to human germlines, we stratified these by the particular human V-region genes. In Tables 4 and 5 we report the most common V-region genes to which the medicinal patent sequences align (corresponding numbers for all patents can be found in Supplementary Table 3). We compare the distribution of germline genes in patents to the germline usage in therapeutic antibodies to show to what extent patent submissions reflect current therapeutics.

The top heavy and light V region genes are identical among medicinal patented sequences, medicinal patents and therapeutics. The most used human heavy chain V-gene by sequence, family, and therapeutic usage is IGHV3-23, accounting for 25.3% of all patented medicinal sequences, occurs in 15.6% of all medicinal families and accounts for 16.4% of therapeutics. The most frequently observed human light chain germline usage is IGKV1-39, accounting for 14.6% of all patented medicinal sequences, 12.8% of all medicinal patent families and 18.4% of therapeutic antibodies. Some of the most commonly observed genes might be the result of specific platform choices that might attempt to recapitulate naturally observed frequencies or focus on a small set of scaffolds. The most frequently used germlines broadly correspond between patented sequences, medicinal patents and therapeutics, even though the ordering might not be the same. This indicates that the patent literature well reflects the

**Table 5.** Top 20 most common human V-region genes antibodies from patents aligned to. For each patent antibody sequence for medicinal applications (MedPatAb) that aligned to human germline V-regions, we noted the IMGT V-region gene. We show the number of unique sequences that aligned to a given human V-region gene (Per Sequence) and number of patent families these originated from (Per Family). We also show the number of therapeutic antibody sequences in clinical use that align to the given V-region gene (Per Therapeutic).
choices of V-region genes of therapeutic antibodies in clinical use.

**Antibodies from patent documents well reflect therapeutic antibody sequences, with the exception of CDR-H3 lengths**

The germline gene distribution of antibody sequences from patents appears to reflect the germline gene distribution of therapeutic sequences, though such comparison is not fit to indicate the actual sequence discrepancies between the two datasets. We assessed the extent to which patented sequences are a reflection of therapeutics by pairwise sequence comparisons between the two datasets.

For each of the 563 therapeutics, we checked whether we can find a perfect length-matched hit in PAD. For 546 (97%) of 563 therapeutics, we found a perfect length-matched hit in PAD. For the remaining 17 therapeutics without perfect matches, we found that the PAD version used for this study (January 2020) was out of date or high sequence identity matches existed, but not perfect ones. We assessed whether the imperfect matches were caused by certain sequences missing from our database, by comparison to the leading provider of free biological sequences search, Lens.org. Nevertheless, upon searching Lens.org we still could not identify perfect matches to perfect sequences (Supplementary Table 4).

For each antibody sequence from a patent, we noted the highest IMGT sequence identity to any therapeutic and present the results stratified by AllPatAb and MedPatAb sequences in **Figure 3**. A large proportion of PAD sequences align with high sequence identity to one of the 563 therapeutics. A total of 21,772 (16.1%) of heavy chain AllPatAb sequences and 17,378 (18.7%) of heavy chain MedPatAb sequences have matches of 90% sequence identity or better to a therapeutic sequence. A total of 44,919 (40.9%) of light chain AllPatAb sequences and 31,241 (46.2%) of light chain MedPatAb sequences have matches of 90% sequence identity or better to a therapeutic sequence. Altogether, these results illustrate that many sequences in patent documents well reflect the therapeutic antibody sequences currently in clinical use. However, a large number of sequences have matches below 90% sequence identity to either heavy or light therapeutic heavy chain. This could reflect sequences that are only currently in development or never found their way to the clinic as a result of failure or abandonment.

Perfect matches between full variable region PAD sequences and therapeutics implicitly indicates good correspondence in the complementarity-determining region (CDR). Arguably, the most diverse and thus the most engineered portion of an antibody is its heavy chain CDR3, CDR-H3. The length of CDR-H3 has been previously shown to be a good estimator of overall developability of an antibody, with therapeutic antibodies having shorter CDR-H3. We contrasted the CDR-H3 lengths found in PAD to those in therapeutic, structural, and natural human antibodies. We extracted CDR-H3s from antibody structures found in the Protein Data Bank that are regularly collected by the Structural Antibody Database (Saab). The natural human antibodies were sourced from a deep next-generation sequencing study by Briney et al. downloaded from the Observed Antibody Space database. We found a total of 58,383 unique CDR-H3s in all PAD sequences (AllPatAb), 37,247 unique CDR-H3s in antibodies from medicinal patents (MedPatAb), 422 unique CDR-H3s in therapeutics, 2021 unique CDR-H3s in structures, and 73,217,582 unique CDR-H3s from natural human antibodies. We contrasted the distribution of lengths between patents and therapeutics in **Figure 4** and for each of these datasets in Supplementary Figure 2.

The distribution of CDR-H3 lengths from patent sequences does not appear to be different between AllPatAb and MedPatAb sequences. Therapeutic CDR-H3s and structures have the shortest lengths (median 13, mode 12), followed by structures (median 14, mode 12) patents (median 14, mode 13), and natural human antibodies (median 16, mode 15). The shorter lengths in structures might be reflective of the large number of artificial/therapeutic antibodies that can be found in AraB. Lengths of CDR-H3s from patent sequences appear to be mid-range between therapeutic and natural antibodies. This suggests that patent antibody sequences might reflect a certain amount of engineering of these molecules, as they do not follow the natural distribution, normally favoring longer lengths. Nevertheless, patent antibody CDR-H3 do not recapitulate the therapeutic CDR-H3 length distribution. Since most

**Figure 3.** Closest matches of antibody sequences from patents to therapeutic antibodies. For each sequence in AllPatAb and AllPatMed we noted the closest IMGT sequence identity to a therapeutic antibody. (a) Distribution of heavy chain sequence identities to closest therapeutic heavy chain. (b) Distribution of light chain sequence identities to closest therapeutic light chain.
therapeutic CDRs can be found in sequences from patent documents, the discrepancy with the therapeutic length distribution suggests certain engineering choices made for those molecules that are not revealed in this study.

**Patent landscape of single domain antibodies**

Our earlier results revealed that the majority of antibodies from patents align well to human or mouse germline V region genes, which recapitulates the widespread use of the ‘traditional’ antibody format containing both heavy and light chains. The third most commonly identified organism was alpaca (Table 4), which suggests the single domain antibody (sdAb) format. The single domain antibodies are found naturally in camelids (camels, llamas, alpacas), and because of the lack of light chain are believed to have more favorable biophysical properties than antibodies, without detriment to their antigen recognition ability. They have been commercialized as therapeutics by Ablynx under the trademarked name Nanobody®, with the first single domain antibody drug, caplacizumab (Cablivi®), recently approved. The marketing approval of the first sdAb drug holds the promise of more molecules in this format in the near future, which can be reflected by patents.

We identified the total number of patent families in PAD having sdAbs to quantify the possible number of molecules in this format in development, providing an orthogonal view to currently known therapeutic candidates. Patent families were classified as containing sdAbs if they were classified as C07K2317/569 (single domain, e.g., dAb, sdAb, VHH, variable new antigen receptor (VNAR) or nanobody®) or C07K2317/22 (from camelids, e.g., camel, llama, dromedary) or if they contained sequences aligning to alpaca sdAb germlines. Using the classification method, we identified 845 families, and using the alpaca germline method we found 867 families. There was an overlap between the two, resulting in total of 1,176 families identified as containing sdAbs (7.1% of all of our 16,526 families in PAD). Of the 1,176 families, 586 (49.8%) were classified as containing antibodies for therapeutic purposes.

The top 30 organizations sorted by the number of families containing sdAb sequences (Supplementary Table 5) well reflect the companies developing biotherapeutics in this format. The list, however, contains more organizations than those currently reported as developing sdAb therapies, suggesting that the field might be more nuanced, notwithstanding wide use of sdAbs for imaging and diagnostic purposes. From the list of known sdAb therapeutics, our list does not contain AdAlta and Ossianix. AdAlta focuses on production of i-bodies that are proteins engineered to resemble the shape of shark antibodies, whereas Ossianix specializes in VNAR molecules, which are natural heavy-chain only antibodies derived from sharks. Neither format is included in our pipeline. In fact, not all sdAbs that we identified follow the natural camelid format, as there exist sequences of single domain human antibodies (e.g., US2011097339).

We checked the total number of sequences in PAD that could be identified as sdAbs. The 1,176 patent families that we identified as containing sdAbs include a total of 48,849 unique heavy chain sequences. Not all such sequences are sdAbs, as the patent document might have included traditional antibodies as well. Therefore, we calculated the number of sequences that were identified as alpaca sdAb germlines and sequences found in one of the 1,176 families, but containing only heavy chains. We found a total of 12,914 unique sequences aligning to sdAb alpaca germlines and 13,368 unique sequences found in 1,176 sdAb families containing heavy chains only. There was an overlap between the two sequence sets and combining them resulted in a total of 15,792 possible sdAb sequences, which

---

**Figure 4.** Distribution of CDR-H3 lengths. We plotted the distribution of CDR-H3 lengths from therapeutic antibodies (Antibody Therapeutics) and all antibodies from patents in PAD (Patents, AllPatAb).
makes up 11.7% of all the 135,397 heavy chain sequences in PAD. Of the 15,792 possible sdAb sequences, 8,342 (52.8%) were found in patent documents classified as containing antibodies for medicinal purposes. Therefore, single domain antibody sequences appear to make up a non-trivial proportion of antibody sequences found in patents, which could be indicative of upcoming sdAb clinical trials and approvals.

In order to provide an indication of the possible activity to come in the field of single domain antibodies, we plotted the earliest and most recent dates associated with any of the 586 sdAb patent families classified as having antibodies for medicinal applications (Figure 5). There appears to be a steady increase in the number of patent documents including sdAb sequences for medicinal purposes, which also holds true for all 1,176 patent families containing sdAb sequences (Supplementary Figure 3). Given the steady rise in the number of patents containing sdAbs and recent approval of caplacizumab, one might expect more molecules in this format in clinical use in the future.

**Engineering perspective of therapeutic antibody properties**

In this work we established that the majority of antibody sequences in patent documents are destined for medicinal applications. Furthermore, we showed that primary therapeutic sequences can indeed be found in patent submissions, suggesting that engineering information encapsulated in such successful molecules is contained within patent submissions. This opens opportunities for mining such engineering knowledge in the future.

To provide a reference point to information that could be exploited from patent documents for therapeutic antibody development, we identified the properties associated with antibodies in patents. We listed the most commonly cited CPC patent classes that accounted for more than 2% of all of our patent families in Supplementary Table 6. This provided insight into properties that are most commonly reported, such as humanization (C07K2317/24, 27.9% of patent families), scFvs (C07K2317/622, 17.7% of patent families), or multispecificity (C07K2317/31, 11.5% of patent families).

We studied Supplementary Table 6 to compile an extract of certain antibody engineering properties that we present in Table 6. We identified properties relating to humanization, scFvs, Fc engineering, multispecific antibodies, fragments of antibodies, and fusion polypeptides. We also included an overarching ‘general’ group that aimed to capture more general properties, such as stability and cross-reactivity. For each of the groups, we identified CPC classes that we considered as related to these groups, supplemented by certain keywords (e.g., bispecific, trispecific). We then calculated the proportion of patent documents that matched either the class or keywords of a given group.

The groups in Table 6 exemplify a constrained set of documents that should encapsulate different engineering properties of antibodies. We propose that such extracts should be a good starting point for attempts to extract information potentially benefiting design of particular antibody properties. For instance, sequences contained within the humanization grouping could capture modifications introduced to non-human sequences aiming to deimmunize these. These could provide a more precise insight into decisions and mutations introduced, beyond simple germlining or distances from non-human sequences that are currently used for computer-based humanization. Similarly, studying sequences from patents that reported stability improvement could provide engineered modifications that are aimed to improve manufacturability properties.

Therefore, groups presented in Table 6 offer an indicator of the opportunities for exploiting the engineering knowledge contained within patent documents. Nevertheless, in each case we envisage that a bespoke approach would have to be
Table 6. Volume of patent families corresponding to different engineering properties of antibodies. Each engineering property of antibodies is associated with CPC patent classes and keywords.

| Property                  | #Patent families (%) | keywords | classes |
|---------------------------|----------------------|----------|---------|
| HUMANIZATION              | 28%                  | Humaniz* | A01K2207/15:Humanized animals |
|                           |                      |          | C07K2317/24:Containing regions, domains or residues from different species, e.g. chimeric, humanized or veneered |
| GENERAL                   | 24%                  | n/a      | C07K2317/35:Valency |
|                           |                      |          | C07K2317/39:Stability, e.g. half-life, pH, temperature or enzyme-resistance |
|                           |                      |          | C07K2317/41:Glycosylation, sialylation, or fucosylation |
|                           |                      |          | C07K2317/72:Antibody-dependent cellular cytotoxicity [ADCC] |
|                           |                      |          | C07K2317/74:Complement-dependent cytotoxicity [CDC] |
|                           |                      |          | C07K2317/33:Crossreactivity, e.g. for species or epitope, or lack of said crossreactivity |
| SCFV                      | 17%                  | scfv, single chain variable | C07K2317/622:Single chain antibody (scFv) |
| FC-ENGINEERING            | 15%                  | Fc,Fragment crystallizable | C07K2317/72:Increased effector function due to an Fc-modification |
|                           |                      |          | C07K2317/52:Constant Fc region, isotype |
|                           |                      |          | C07K2319/30:Non-immunoglobulin-derived peptide or protein having an immunoglobulin constant or Fc region, or a fragment thereof, attached thereto |
|                           |                      |          | C07K2317/526:CH3 domain |
|                           |                      |          | C07K2317/64:Comprising a combination of variable region and constant region components |
|                           |                      |          | C07K2317/53:Hinge |
|                           |                      |          | C07K16/468:Immunoglobulins having two or more different antigen binding sites, e.g. multifunctional antibodies |
| MULTISPECIFICS            | 13%                  | Bispecific, trispecific | C07K2317/31:Multispecific |
| FRAGMENTS                 | 12%                  | Fab,Fab' | C07K2317/55:Fab or Fab' |
| FUSION POLYPEPTIDE        | 10%                  | Fusion polypeptide | C07K2319/00:Fusion polypeptide |

used, that is nonetheless greatly facilitated by our effort to extract patent antibody sequences and their related metadata.

Discussion

Successful exploitation of antibodies as therapeutics relies on ever deeper understanding of the biology of these molecules. Many features of therapeutic antibodies can be found in naturally sourced sequences, but effective biotherapeutic requires bespoke engineering for clinical safety and developability. We proposed that such biotherapeutic engineering knowledge could be reflected in patent documents containing antibody sequences.

Our analysis of patents containing antibody sequences revealed that the majority of such documents are explicitly developed for therapeutic antibodies. Most therapeutic antibody sequences can be found in patent documents. Further to that, many sequences from patents are within close sequence identity of therapeutic antibodies that are approved or undergoing clinical trials. This suggests that thousands of antibodies from patents could provide a reflection of engineering choices that were made during the development of therapeutic molecules. Such data could offer an integrated collection of insights into the features that were designed into antibodies to make them successful therapeutics.

This information could be readily exploited by computational methods. Raybould et al. have previously shown that analysis of only 137 CST antibodies can provide insights into developability of these molecules. As demonstrated by our analysis, there is an order of magnitude more patented sequences that are close sequence matches to such CSTs. These could indicate different variants and possible features of biotherapeutics, creating a more complete picture of what makes a successful biotherapeutic. Studying subsets of the patented sequences, e.g., stability-engineered antibodies, could provide commonalities shared between such modified antibodies that could be exploited in the future. Contrasting the patented and naturally sourced sequences by features, such as gene usage or sequence identities, might provide insights into developability of therapeutics. For instance, naturally sourced sequences can offer an indication of self-tolerated features that might be absent in therapeutic antibodies, increasing their immunogenicity. The patent sequence data available now could provide a sound basis to tackle such large-scale comparisons.

The use of antibody sequences from patents, however, is not without its caveats. For many years, patent sequence deposition was not standardized, and to this day the data are fragmented, hampering comprehensive collection efforts. Certain commercial patent sequence data providers such as Clarivate’s GENEOSEQ® collect biological sequences with support from manual curation efforts to ensure they capture all the information across heterogeneous depositions. Furthermore, unlike academic literature, patent documents are designed to provide legal protection. This might result in wide claims on sequence identities to proposed antibody variants that could obfuscate the resulting therapeutic sequence. As we demonstrated, antibodies in patents provide a good reflection of the therapeutics either approved or in clinical use. This would suggest that even though claims could be quite wide on sequence space, many of them appear to fall within the sequence identity orbit of currently available therapeutics. Therefore, certain antibody sequences from patents could broadly reflect the engineering choices in the design of these molecules.

The already large number of antibodies from patent documents will most likely keep rising, as we demonstrated by the growth in the number of such documents in the recent years. In fact studying such patents could provide an early indication of approvals to come. This might be specifically true in the sphere of single domain antibodies. There is just one such approved therapeutic on the market and 10 in clinical trials (in 2019). We find a great number of sdAb patents,
suggesting that the field might further develop in the near future, providing an alternative to traditional (i.e., full-length) monoclonal antibodies therapeutics.

The ongoing increase of patents describing antibodies for medicinal indications will continue to contribute to an already extensive body of knowledge of antibody engineering. This data could be used to offer insights into the engineering choices in designing these molecules, accelerating delivery of biotherapeutics to the clinic.

Materials and Methods

Identifying antibody sequences in patent documents

Raw biological sequence data associated with patent documents were downloaded from four freely available accessible services: the USPTO (https://www.uspto.gov/), DDBJ,\textsuperscript{12} EBI,\textsuperscript{13} and WIPO (https://www.wipo.int/).

The USPTO data were divided between the full text submissions (https://bulkdata.uspto.gov/) and lengthy sequence listings (http://seqdata.uspto.gov/). Using a custom Python script, the USPTO full text submissions were scanned for nucleotide or amino acid sequences and listings containing these, whereas the USPTO Publication Site for Issued and Published Sequences (PSIPS) contained sequence listings only. Using a custom Python script, the WIPO FTP documents (ftp://ftp.wipo.int/pub/published_pct_sequences) were scanned for nucleotide and amino acid sequences. In both cases of USPTO and WIPO, differences in sequence listing formats from different time periods was accounted for by developing a custom Python parser for each case, transferring all the raw sequences and their associated patent numbers into FASTA format. Data from DDBJ and EBI are available through their ftp services (ftp://ftp.ddbj.nig.ac.jp/ddbj_database and ftp://ftp.ebi.ac.uk/pub/databases, respectively) and were readily available in FASTA format.

The nucleotide entries were scanned for antibody sequence by using IGBLAST\textsuperscript{16} as described previously,\textsuperscript{17} and their amino acid translations were noted. The raw amino acid sequences were scanned for the presence of antibodies using ANARCI.\textsuperscript{18} We only kept those amino acid sequences where all three CDRs and all four framework regions could be identified and that contained only 20 canonical amino acids. This resulted in a dataset of IMGT-numbered amino acid sequences, associated with their patent numbers.

Patent metadata acquisition and antibody target identification

Different patent numbers can point to the same document, submitted across several jurisdictions, termed ‘patent family’. For each patent number associated with a sequence we identified the patent family by using the Open Patent Services API v. 3.2 (developers.epo.org/ops-v3-2). Using the Open Patent Services API, we downloaded the metadata associated with each family, which included: family identifier, title, description, patent numbers with associated dates and applicants.

The patent metadata was used for antibody target identification. Even though there exist certain CPC classifications indicating what the antibody should bind to, we noted that they were not universally present. Therefore, we performed manual target annotation, supported by Named Entity Recognition (NER). We applied the GENIA NER\textsuperscript{38} parser to the titles and abstracts of patent families. As with scientific publications, titles and abstracts can be expected to reflect the most important content of the document,\textsuperscript{39} in particular pertaining to the binding mode of the reported antibody. The resulting annotations accelerated the manual process of annotating each of our patent families with possible targets.

Web Service

Our data is accessible for academic noncommercial use via a web service available at http://naturalantibody.com/pad. Users can search for antibody sequences by pasting the amino acids of the variable domains. The input sequence is IMGT-numbered. The sequences in PAD are IMGT-aligned to the input sequence and the top 50 best sequence identity matches are displayed.

Abbreviations

| Abbreviation | Full Form |
|--------------|-----------|
| CDR          | Complementarity-Determining Region |
| CST          | Clinical Stage Therapeutic |
| DDBJ         | DNA Data Bank of Japan |
| EBI          | European Bioinformatics Institute |
| PAD          | Patented Antibody Database |
| PCT          | Patent Cooperation Treaty |
| PDB          | Protein Data Bank |
| WIPO         | World Intellectual Property Organization |
| USPTO        | United States Patent and Trademark Office |

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

ORCID

Konrad Krawczyk http://orcid.org/0000-0003-0697-5522

References

1. Kaplan H, Muralidharan M, Schneider Z, Reichert JM. Anti- bodies to watch in 2020. MABS. 2020. doi:10.1080/19420862.2019.1703531.
2. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y, et al. Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci USA. 2017. doi:10.1073/pnas.1614608114.
3. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng S, Reddy ST. Deep learning enables therapeutic antibody optimization in mammalian cells. bioRxiv. 2019. doi:10.1101/617860.
4. Norman RA, Ambrosetti F, Bonvin AMJJ, Colwell LJ, Kelm S, Kumar S, Krawczyk K. Computational approaches to therapeutic antibody design: established methods and emerging trends. Brief Bioinform. 2019. doi:10.1093/bib/bbz095.
5. Krawczyk K, Raybould MIJ, Kovaltsuk A, Deane CM. Looking for therapeutic antibodies in next-generation sequencing repositories. MABS. 2019. doi:10.1080/19420862.2019.1633884.
6. Raybould MIJ, Marks C, Krawczyk K, Taddei B, Nowak J, Lewis AP, Bujoztez A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proc Natl Acad Sci USA. 2019. doi:10.1073/pnas.1810576116.
