SCIENTIFIC REPORTS

OPEN

Humanization of Antibodies using a Statistical Inference Approach

Alejandro Clavero-Álvarez[1], Tomas Di Mambro[2], Sergio Perez-Gaviro[3,4,1], Mauro Magnani[2] & Pierpaolo Bruscolini[4,1]

Received: 2 May 2018
Accepted: 19 September 2018
Published online: 04 October 2018

Antibody humanization is a key step in the preclinical phase of the development of therapeutic antibodies, originally developed and tested in non-human models (most typically, in mouse). The standard technique of Complementarity-Determining Regions (CDR) grafting into human Framework Regions of germline sequences has some important drawbacks, in that the resulting sequences often need further back-mutations to ensure functionality and/or stability. Here we propose a new method to characterize the statistical distribution of the sequences of the variable regions of human antibodies, that takes into account phenotypical correlations between pairs of residues, both within and between chains. We define a "humanness score" of a sequence, comparing its performance in distinguishing human from murine sequences, with that of some alternative scores in the literature. We also compare the score with the experimental immunogenicity of clinically used antibodies. Finally, we use the humanness score as an optimization function and perform a search in the sequence space, starting from different murine sequences and keeping the CDR regions unchanged. Our results show that our humanness score outperforms other methods in sequence classification, and the optimization protocol is able to generate humanized sequences that are recognized as human by standard homology modelling tools.

Antibody-based drugs have acquired an increasing importance in the last two decades, both for imaging and for therapeutic uses, especially to treat different types of cancer and autoimmune diseases. However, their development is a long and difficult process, prone to fail at different stages. Antibody humanization is a key step in this process, unless the candidate is already obtained from a human library, and is essential in moving from the preclinical to clinical stage. In fact, new antibodies are typically developed in animal models (most often, in mouse); however, the antibodies obtained by this way are usually not tolerated by humans, eliciting *in vivo* an immune response against the murine antibody. Thus, they need to be "humanized", substituting part of their sequences by the human ones, while preserving their specificity, affinity and stability. Although computational methods are available, nowadays such humanization process is mostly a trial-and-error process, based on CDR-grafting and back mutations[1]. CDR-grafting implies selecting the Complementarity-Determining Regions (CDRs), responsible for antigen recognition, from the given murine sequence and grafting them into the human Framework Region (FR); the latter is selected by looking, in the human genome, at the germlines that produce FRs most homologous to the murine ones: the hope is that the combination of such human frameworks with the original murine CDR will result in a molecule that still preserves its stability and activity, but is tolerated by the human immune system. However, most of the times this approach is not completely successful at either of its quests, and the researcher is left alone in trying further mutations, until an antibody with the selected properties is identified. This error-prone process is a true bottleneck in the development of new treatments, in a market of increasing global impact.

From an algorithmic point of view, CDR-grafting corresponds to a search, in the human germlines, for the sequence with minimal Hamming Distance (i.e highest similarity) to the original murine one. Thus, the similarity to the closest germline represents a "humanness score" to maximize, for this approach. More in general, any humanization protocol will rely on maximizing some "humanness score", whose basic requirement is to be able to distinguish human from mouse sequences with as few errors as possible. Most humanness scores are based on pairwise sequence identity between the sample and a set of reference (most often, germline) human sequences: for instance, the score can correspond to the average similarity[2], or the average among the top 20 sequences[3], or

[1]Departamento de Física Teórica, Universidad de Zaragoza, Zaragoza, 50009, Spain. [2]Department of Biomolecular Sciences, University of Urbino "Carlo Bo", Urbino, Italy. [3]Centro Universitario de la Defensa, Zaragoza, 50090, Spain. [4]Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Zaragoza, 50018, Spain. Correspondence and requests for materials should be addressed to P.B. (email: pier@unizar.es)

the highest similarity, over "windows" of typically 9 residues[4,5]. Recently, a different approach as been proposed by Seeliger[6], introducing a score function that accounts both for local preferences and for pair correlations between residues at different positions. Interestingly, such approach is reasonably capable to distinguish between human and mouse sequences, despite a relevant residual overlap between their distribution; also, the stochastic humanization process under such score function samples regions of low immunogenicity, even though the final basin of attraction of the trajectories presents an intermediate value of immunogenicity (as measured by the Epivax score)[7]. It is also worth mentioning that such approach breaks with the common logic of considering CDRs as the only antigen binding regions, by treating correlations between any positions (CDR or framework) on the same grounds, which can be a safer option due to the fact that there are relevant antigen binding residues also in framework regions[8]. However, Seeliger's approach uses an ad hoc score function for the pairs of residues, $\ln(p_{ij}^2(A_i,\ A_j)/(p_i(A_i)p_j(A_j))$, that apparently is loosely related to mutual information $MI_{ij} = \sum_{A_i,A_j} p_{ij}(A_i,\ A_j)\ \ln(p_{ij}(A_i,\ A_j)/(p_i(A_i)p_j(A_j))$, and may suffer its same problems[9], in distinguishing direct and indirect correlations. A more fundamental approach deals with the observed sequences as instances of a general probability distribution $p(\mathbf{A})$ over the sequences $\mathbf{A} = (A_1,\ \ldots,\ A_L)$. By constraining the sequences according to the observed site $f_i(A_i)$ and pair $f_{i,j}(A_i, A_j)$ frequencies, and looking for the distribution maximizing the Shannon entropy, one finds $p(\mathbf{A}) = e^{-H(\mathbf{A})}/Z$, where: $H(\mathbf{A}) = -\sum_{i<j} e_{ij}(A_i,\ A_j) - \sum_{i=1}^{L} h_i(A_i)$ and $Z = \sum_{\mathbf{A}} e^{-H(\mathbf{A})}$. This is appealing, since it paves the way for a connection to statistical physics: $p(\mathbf{A})$ appears as a Boltzmann distribution, corresponding to the energy function H, made up of a one-body term $h_i(A_i)$, stating the preferences of each position for each amino acid type, plus an interaction term e_{ij}, coupling the position i and j. These effective interactions should be regarded as the expression of the different constraints that intervene in the accelerated sequence evolution that immunoglobulins undergo during maturation, as for instance, the optimization of the interaction with the antigen, the need to preserve stability of the folded structure against unfolding or misfolding and aggregation, the requirement of low affinity for the T-cell receptors, to avoid an immunogenic response. The knowledge of the parameters $e_{ij}(A_i, A_j)$, $h_i(A_i)$ above allows to calculate the probability of any sequence, opening the possibility to associate to each sequence a measure of its "humanness". However, the inference of the parameters is a formidable task, that cannot be accomplished in an exact way. Techniques as Direct Coupling Analysis[9] have been used to provide approximate estimates; here instead, we follow the approach by Baldassi and coworkers[10] specifying a simple, quadratic form for the energy function H, that allows an analytic derivation of the parameters and of the posterior probability of any sequence. This approach has been recently used to predict antigen-antibody affinity[11], with a different choice of the regularization, and using the quadratic form as a sequence score; here we elaborate on the original approach, and derive a score that is related directly to the posterior probability of a sequence, as explained below. We will still refer to this model as the "Multivariate Gaussian" Model (MG), even if the posterior probability score that we calculate and use is actually a multivariate Student distribution. We use this score to assess its efficiency in classifying murine and human sequences, comparing the method with simpler approaches just based on the Hamming Distance (i.e. sequence pairwise difference) between the input sequence and the human or mouse reference database. To obtain the latter, we curated human and murine learning and test database, of matching VH and VL chains, in order to assess the joint role of the light and heavy region in determining humanness.

Our goal is to understand the role of correlations between mutations at different positions, including across heavy and light chains, as accounted for by the statistical model, in determining the traits of humanness. Also, we want to understand whether the VH and VL chains behave and vary independently, as assumed implicitely by all humanization methods that deal separately with the two chains, or if correlations between the VH and VL chains play a relevant role, and cannot be neglected. Moreover we consider the relationship between our humanness score and the observed immunogenicity of a small set of antibodies for which the immunogenic reaction in patients has been reported in the literature. Finally, resorting to Steepest-Descent/Monte-Carlo simulations with the MG statistical score, we perform the humanization of a few murine sequences that have also been humanized experimentally, to see how our results compare with the experimental ones. We conclude by discussing the possible applications and future lines of research stemming from this approach. Before proceeding, let us mention that alternative strategies, to define a humanness score based on sequence distributions, could be attempted: for instance, one could resort to Hidden Markov Model (HMM) techniques to learn a probability profile from the alignment of the sequence database. However, such an approach does not appear to be very appealing, for our goals: indeed, Seeliger's results[6] suggest that correlations between pairs of residues are relevant. HMMs are effective in accounting local correlations between matching columns or neighboring sites (i.e. correlations that can be encompassed by the transition matrix from the state at position i to that at position $i+1$ in the model graph), but are not suitable to describe generic long range correlations[12], as those between VH and VL, whose relevance we want to assess. This does not rule out that alternative scores based on HMM could be effectively introduced, but we do not adopt this strategy here, and leave the subject to future investigations.

Results

Correlated and uncorrelated classification of test databases.

After creating human and murine, learning and test databases of matching VH, VL regions, and fixing the parameters of the statistical model as described in Methods, we compare the performance of the distance based approaches with the statistical distribution method (Fig. 1), in distinguishing human from non-human (in our case, murine) sequences in the test database. Our goal is to assess if and to what extent the MG approach, accounting for two-sites correlation, improves the classification based on sequence-similarity criteria. We can see that the MG model efficiently distinguishes between murine and human sequences in the test database, scoring better than the distance-based approaches. Notice also that correlations do matter: as already pointed out by Asti and coworkers[11] in a similar context, setting to zero the non-local interactions, and keeping only a block-diagonal distribution for the correlations (see

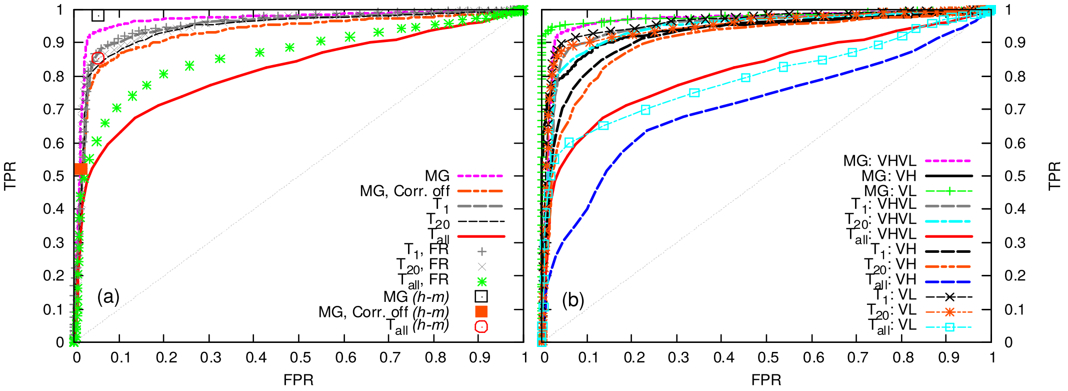

Figure 1. ROC curves for the different models. ROC curves obtained upon classification of the test human and murine databases, using the human learning database to learn the MG distribution, and to calculate the average distance for the "T_{all}" and "T_n" methods. Panel (a): ROC curves obtained using the full VHVL chain for classification, with or without CDR regions. Panel (b): ROC curves obtained using the VH or VL chain separately for classification. "Corr. off" indicates that correlations between residues have been removed in the MG model, "FR" refers to the curve obtained removing the CDRs and keeping just the framework regions, "h-m" indicates that classification is performed using both human and murine learning datasets as reference (see Methods).

Methods), results in a much worse performance. Interestingly, if we just eliminate correlations between the VH and VL regions, keeping those within the same chain, the performance is practically unaffected (data not shown): this supports the commonly used approach to deal with heavy and light chains separately.

Distance-based methods similar to the one proposed by Adler[3], considering only the distance to the n closest sequences to the query one in the learning database, provides a better classification than accounting for the distance from the whole set of human learning sequences[2] (the "T_{all}" method). In fact, the ROC curves steadily improve as we move from T_{all} to T_1, suggesting that the local structure of the sequence space may be more relevant than the global one, and that the heterogeneity in the human database can be misleading for a distance-based classification.

As it could be expected, upon removing the CDRs and keeping only the framework regions, the predictions by T_n improve (especially, for increasing n), supporting that the CDRs do not carry any relevant species-specific information. For the sake of completeness, it is also interesting to notice that using two different reference learning databases improves the classification performance of the MG model, but much more so that of the T_{all} approach: classifying a sequence according to whether its distance score to the human learning dataset is better than that from the murine learning dataset is very fast and efficient. However, this approach cannot be safely adopted for humanization, where the goal is to make a sequence "sufficiently human", and not necessarily "as close to human and far from murine as possible".

Classification based on just the VH or VL regions suggests that the VL region allows to better distinguish human from murine sequences: for the MG case, VL-based classification performs even better than the VHVL-based one, while VH based classification does not reach the same results. The same is true also for the T_n method. Thus, apparently the VL carries a greater amount of information on the human or murine nature of the sequence. Table S3 in Supplementary Information (SI) reports single-value indicators of the quality of the classification of the test databases with the different methods, complementing the information reported in Fig. 1.

Classification of test, murine, chimeric, humanized and fully human antibodies. We classify the test database and the engineered antibodies according to the methods above, using the VHVL distribution, since this is the one that we will use later on for sequence design. We use the human learning distribution as the reference one for classification, and we fix the threshold score, separating human sequences from the rest, according to the value that maximizes the Youden's index for the test database (see Methods; in general, the values maximizing the latter and Matthews Correlation Coefficient almost coincide, ensuring that the choice of a particular indicator is not crucial).

Figure 2 reports the distributions of the scores for the learning and test databases, as well as for the pharmaceutic antibodies whose sequences are publicly available[13–15] (see Methods and SI Table S1), labeled by the suffix of their International Nonproprietary Name (INN). Notice that the MG-scores of the human learning datasets are much higher than those of the human test dataset (and actually quite close to the maximum possible, $MG_{max} = 13124.49$). This is most likely a signal of overfitting of the learning dataset, which is inevitable due to the dimension of the matrix Σ and the number of aligned sequences. We expect that, as new researches allow to increase the size of the learning database, this issue would become less evident. Table 1 reports the fraction of correct predictions (for the therapeutic antibodies, we consider correct a prediction of -umabs and -zumabs as human, -omabs and -ximabs as murine).

It can be noticed that MG yields the best results, with the highest fraction of correct classifications for the human and murine test databases; the performance of all the method in classifying therapeutic antibodies is

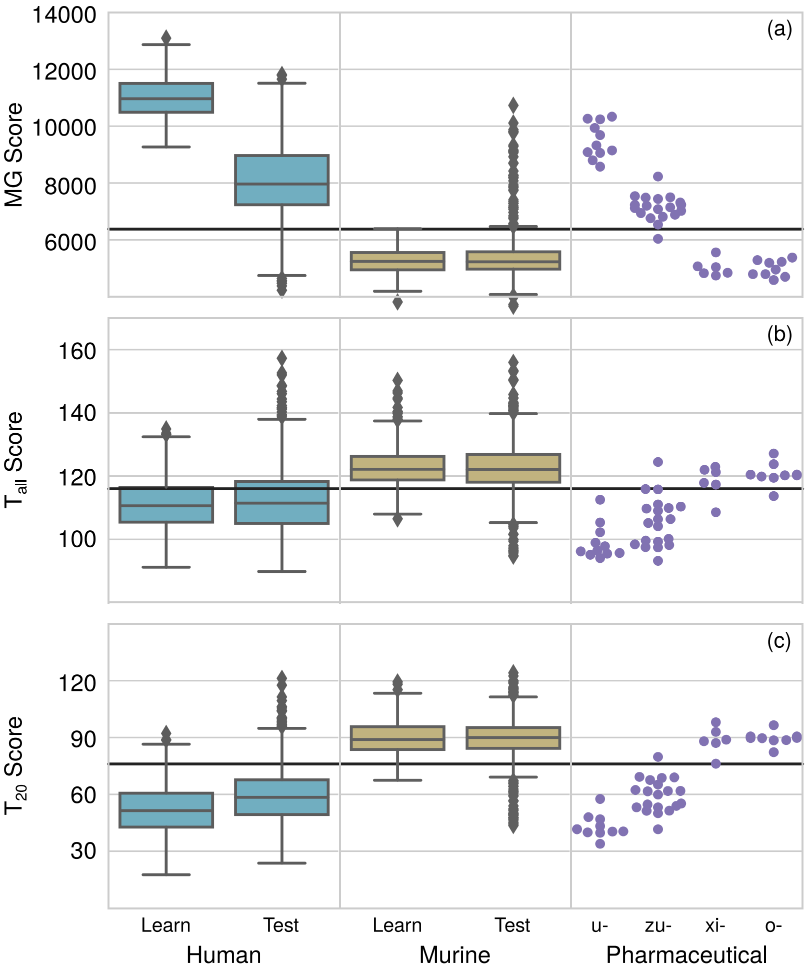

Figure 2. Boxplot of three score distributions for the different datasets: Panel (a): MG score; Panel (b): T_{all} score; Panel (c): T_{20} score. Pharmaceutical antibodies are indicated according to the suffix in their International Nonproprietary Name: "umab" are fully human antibodies; "zumab" are humanized antibodies, usually containing murine CDRs grafted on top of human framework variable regions; "ximab" are chimeric antibodies, obtained by assembling the whole murine variable region on top of a human constant part; "omab" are murine antibodies. Since we deal with just the antibodies' variable regions, "ximab" and "omab" are indistinguishable. The horizontal lines signal the threshold score above (for the MG case) or below (for the other cases) which the sequence is classified as human. The threshold values are $t_{MG} = 6383, t_{T_{all}} = 116, t_{T_{20}} = 76$.

	Human	Murine	umab	zumab	ximab	omab
MG	1289/1388	1324/1379	11/11	19/20	6/6	9/9
T_1	1225/1388	1296/1379	11/11	19/20	6/6	9/9
T_{20}	1195/1388	1274/1379	11/11	19/20	6/6	9/9
T_{all}	935/1388	1193/1379	11/11	19/20	5/6	8/9

Table 1. Fraction of correct predictions. Fraction of correct predictions for the test and therapeutic databases, using the threshold obtained as specified in Methods, to distinguish between human and murine sequences.

similar, and almost perfect, for all methods, even if in that case the number of sequences is too small to be statistically relevant.

Correlation of the MG score with immunogenicity. A good performance in classification tasks is the fundamental requirement for a humanness score, but it could be not enough if we want to use such score for antibody humanization. The reason is that sequences with better score might not be better sequences, since the MG score has been derived from sequence information only, and it does not explicitly account for e.g. protein stability, solubility, or immunogenicity. In particular, the latter is a crucial aspect for humanization, since it is highly desirable that a drug does not elicit an immune response. In the following we focus on the MG-score, as the best candidate to use in a humanization protocol. In order to estimate how good is the MG humanness score for humanization tasks, we compare it with the reported fraction of observed immunogenic response (as measured

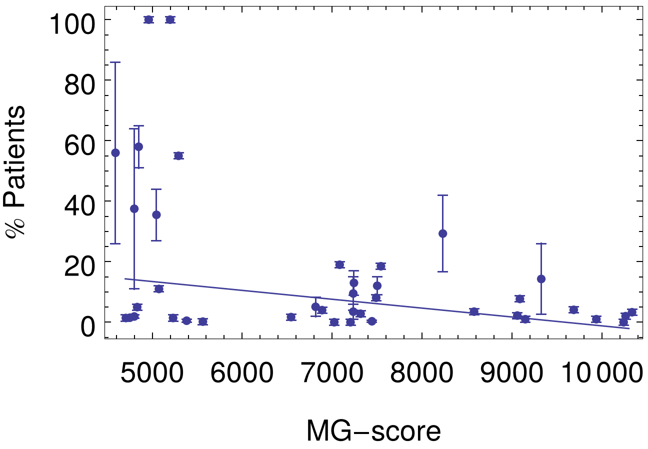

Figure 3. Scatter plot of the experimental immunogenicity and the MG score. The immunogenicity (% of patients that develop antibodies against the therapeutic antibody) is plotted versus the MG-score.

by the appearance of anti-drug antibodies) for several pharmaceutical sequences: Fig. 3 reports the correlations between the MG score and the experimental findings. Notably, clinical essays often report several values for the same drug, since immunogenicity appears to depend on the disease the drug is used for, as well as on the possible combination of the drug with other treatments. A little arbitrarily we have used the mean between minimal and maximal reported values as the average, and the semidifference between them as the error bar of each "measure" in the plot (when just one value is reported, an error of 1% is assumed). Then we have performed a linear fit of the experimental immunogenicity (with uniform weights), reported in the figure. As it can be seen, the experimental values are widely spread, especially at the murine end; this implies that immunogenicity is just loosely related to the degree of humanness of the variable regions. Hence, as expected, the fit $y = 56.35 - 0.005806\,x$ (with x the MG-score and y the experimental immunogenicity) is quite bad, with a low value of the "explained variance" $R^2 = 0.18$, and a Pearson correlation coefficient of $C = -0.43$. However, the fit reveals a global trend of improvement in immunogenicity as the humanness score increases from murine to fully human antibodies.

The limited number of sequences for which an experimental estimation of immunogenicity exists, made us look for an alternative, in-silico definition, that we could apply to any proposed sequence. To this end, we have built an immunogenic score resorting to the MHCII software[16,17] as described in SI Methods, predicting T-cell receptor epitopes in any given sequence, as a proxy for its immunogenicity: our goal was to compare the MG humanness with the immunogenic score for a broader set of sequences, and to test for both scores during the humanization process. However, the results, reported in SI Figs S2 and S3 do not indicate that such immunogenic score is sufficiently reliable, since its correlation with experimental data is even worse than that of the MG-score. Thus, we did not proceed further on this line.

Statistical optimization of murine sequences. Next, we choose in the literature seven pairs of murine-humanized sequences (SI Table S2) and perform Steepest Descent (SD) and Simulated Annealing Monte Carlo (SAMC) simulations starting from the murine sequences, to obtain the sequences with the best MG-score. In all cases, we keep the CDRs fixed, since thet are associated to the antigen-recognition function that we want to preserve. We use two different algorithms due to their different nature: in the SD protocol, at each step we make the point mutation that most improves the score. The output is the closest local optimum to the original murine sequence. On the other hand, the SAMC procedure allows a more extensive exploration of the sequence space, and is less prone to get trapped in local minima, while looking for the global optimum.

Figure 4 reports the MG score versus the Hamming distance to the original murine sequence, starting from each of the murine targets in SI Table S2, while a typical time-course for SD and SAMC simulations is reported in SI Fig. S4 for the first murine target. SI Fig. S4 shows that the MG score steadily increases with time in SD simulations, and the distance to the original murine sequence increases as well, which reinforces our previous findings that the MG score is a good measure of "humanness". This correlation between time and HD from the original sequence results in monotonic curves for SD simulations in Fig. 4, from the original murine sequence on the left, to the final humanized one on the right. On the other hand, the distance to the experimentally humanized sequence in SI Fig. S4 does not present a monotonic decrease and does not reach zero, as expected: the experimentally humanized sequence need not be the only possible one, or even the "most humanlike" one at given CDRs. However, it is interesting to notice that, for all sequences, many of the proposed mutations coincide with those appearing in the experimentally humanized sequence (see column 7 in Table 2). In general, the SD final sequences contain more mutations than the experimental humanizations, suggesting that it is probably unnecessary to fully optimize the MG score, and that viable sequences can be found earlier in the optimization process, whose last steps involve usually smaller changes in the score. Notably, in all cases the MG-score of the experimentally humanized sequence is higher than the original murine one, again confirming the goodness of the MG-score as a humanness score; moreover, the score of the optimal SD sequence is higher than that of the experimental sequence. Another interesting comparison between SD predicted and experimental humanized sequences is shown in Table S4 in SI, where we report the result of a protein-BLAST search for the most similar

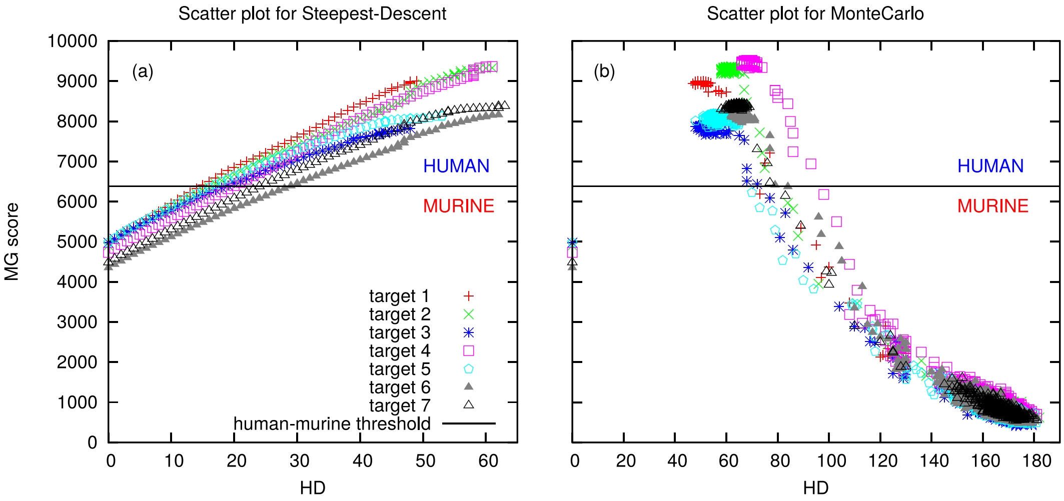

Figure 4. Scatter *MG score-Hamming distance* plot for the Steepest Descent (Panel (a)) and the SAMC (Panel (b)) for all our targets. Here HD is the Hamming distance between our proposed sequence and the original murine one. The SAMC trajectories start at the murine sequences, on the left (HD = 0, MG-score around 4500), then jump immediately to the region at the bottom-right of high HD, low score (highly non-human sequences, but very different from the original murine). Notice that all trajectories roughly overlap in this region: there is no memory of their different, and fixed, CDR sequences, and we witness a basically free exploration of the sequence space. Then, when the temperature falls below a certain threshold, the trajectories move to the top-left region, of highly human sequences with score and HD depending on the fixed CDR regions of the original sequence.

tgt	MG (o)	MG (h)	MG (p)	HD (p, o)	HD (h, o)	TP	FP	TN	FN	FPR	TPR	YJS
1	4963	7571	9000	49	37	31	18	246	3	0.07	0.84	0.78
2	4905	7184	9328	61	48	43	18	234	3	0.07	0.90	0.83
3	4976	6780	7817	48	37	24	24	241	9	0.10	0.65	0.57
4	4734	5689	9362	61	38	22	39	228	9	0.15	0.58	0.46
5	4978	7202	8131	53	39	37	16	243	2	0.06	0.95	0.89
6	4350	5193	8159	62	29	18	44	232	4	0.16	0.62	0.48
7	4481	5468	8379	63	35	24	39	229	6	0.15	0.69	0.56

Table 2. Comparison between original, simulated and experimentally humanized sequences. Column 2, 3 and 4 report the MG-score for the original murine sequence "o", the experimentally humanized sequence "h", and the one predicted by SD "p", for each target (column 1). Column 5, 6 report the number of mutations between pairs of sequences (HD). We define as "positive" (P) the mutations of the predicted sequence with respect to the murine one: P = HD (p, o); "negative" (N) the number of corresponding identical residues in the predicted and murine sequence. Accordingly, True Positive (TP) will be the number of common mutations, with respect to "o", shared by "p" and "h"; False Positive (FP) indicates that in the predicted sequence there is a mutation with respect to the murine, but such mutation is not the present in "h" (or it is not the same mutation); True Negative (TN) imply that neither the predicted nor the humanized sequence have mutations, while False Negative (FN) indicates that the humanized sequence present a mutation with respect to the murine, but the predicted sequence does not. Schematically, being A, B, C possible aminoacids for the triplet (murine, humanized, predicted), we have: (A, A, A) → TN; (A, A, B), (A, B, C) → FP; (A, B, A) → FN; (A, B, B) → TP; thus, HD (p, h) = FP + FN. The True and False Positive Rates are defined as TPR = TP/(TP + FN); FPR = FP/(TN + FP). The last column is the Youden's index: see Methods.

sequence. Interestingly, the experimentally humanized sequence is not always recognized as human by BLAST, while for the SD sequence, always a human or humanized sequence appears as the most similar one, which again supports the goodness of the MG-score as a humanness score.

SAMC simulations start at high temperature (see Methods), which induces an immediate jump from the original murine score, with HD equal to zero, to very low scores and high distances from the initial murine sequence (see SI Fig. S4): the system explores sequences with basically all the residues mutated with respect to the original murine (since we fix 111 CDR positions, the maximal possible HD is 187). The system accepts changes with probability around 0.9, suggesting that the initial temperature is high enough to cross the "energy barriers" and freely explore the sequence space. During the simulation, the temperature is slowly reduced, up to a point where there is a sudden transition from an entropy-dominated regime, where the system explores many sequences with low score (high energy) to an energy-dominated region, where the system get stuck in a high score (low energy) region of the sequence space, ideally where the optimal sequence is located. Correspondingly, the Hamming distances suffer a sudden change as well, and reduce their fluctuations to a small region around their final value.

In Fig. 4 the course of the SAMC simulation can be followed by observing that all targets' trajectories share an initial jump from HD = 0 to a region of high distance, close to the maximum possible one, and low MG score, and then move towards the final values of high MG-score and intermediate HD. Notice that the trajectories finally explore highly-human regions of several hundreds of sequences; however, there is no sampling of the region between the original murine sequence and the final basin of high-scoring sequences. On the other hand, the SD trajectory explores precisely this region, providing humanized sequences of increasing dissimilarity from the original sequence. Notably, the maximal score sequence obtained by SAMC usually does not coincide with (and, obviously, is better than) the best SD sequence: the difference can be of a few residues, pointing at essentially the same minimum in the sequence space, but also of more than 20 residues (such in the case of target 4), suggesting that in some case the SD evolution can become trapped in a local energy minimum, quite different from the global one. This could appear surprising, given that the energy function is a continuous function in N dimensions, with a unique global minimum; however, our dynamics is forced to be discrete, and we keep the CDR residues fixed: this could induce barriers in an otherwise funnel-like landscape, that can trap the SD trajectory, that is bound to find a local energy minimum, while the SAMC procedure has more chances to escape local traps and find the global optimum. On the other hand, the SD approach is convenient for our goals because it naturally provides, by construction, a trajectory of locally optimal solutions at any given distance from the original murine sequence, that is, the sequence of "best humanizing mutations" to perform, starting from the murine sequence; this is extremely useful within a conservative strategy to humanization, aiming at finding a functional humanized sequence rather than the best scoring one, while keeping the number of mutations as small as possible, in order to avoid affecting the stability or solubility of the original protein. The above results suggest that in a first approach, the SD trajectory could yield sufficient information to obtain reliable humanized sequences, at different similarity to the original murine one; SAMC could be unnecessary, unless the SD sequences prove to be unfit and there is a need to obtain an alternative list of several different high scoring sequences.

Discussion

In this article we have introduced a statistical score, built on the Multivariate Gaussian Modeling[10], that is more reliable, as a measure of the "degree of humanness", than the distance of the query sequence from the ensemble of human sequences, or to a subset of it (the T_n methods). Indeed, we have seen that its performance in distinguishing human from murine sequences is higher than the methods relying on just the sequence similarity, due to the fact that the statistical model, on which the MG-score is based, accounts for pair-correlations between residues at different positions. Such correlations represent a key features for the performance of the model, that indeed drops dramatically if they are neglected. However, neglecting correlations between VH and VL regions does not affect the classification performance; since all our databases are restricted to VH-VL pairs that indeed are part of the same antibody, the irrelevance of correlations between VH and VL regions is not the artefact of a random juxtaposition of the two chains, but points to a very limited role of the interplay between the two regions, at least as far as classification is concerned. If translated to humanization tasks, this observation would support the common approach of humanizing the light and heavy chain independently.

The proposed humanness score shows a correlation with the experimental immunogenicity of therapeutic antibodies, even if such correlation is far from perfect. On the other hand, the comparison is made difficult by a series of factors affecting the experimental data: first, immunogenicity is studied in a clinical environment, measuring the frequency of anti-antibody reactions in sets of patients that necessarily represent inhomogeneous samples, due to their age, physical conditions etc.; on top of this, the same antibody, used for sets of patients affected by different diseases or in combination with other drugs, triggers different immunogenic responses. Also, even murine therapeutic antibodies, approved for long-duration treatment, are very likely to present little immunogenicity (imaging antibodies can have more, due to a more sporadic use), so that they are not expected to present the "typical immunogenicity" of a random murine antibody, and this surely introduces a bias in the data. Unfortunately, there is at present no alternative tool that can reliably predict the immunogenicity of a given antibody, so it was not possible to study the goodness of our humanness score as a immunogenicity predictor, for a sufficiently large dataset, free from the biases of therapeutic antibodies.

In any case, the analysis of murine-humanized antibody pairs reveals that, systematically, the MG-score increases upon humanization, thus indirectly confirming the value of the score for humanization tasks. Starting from the murine sequences, we have performed in-silico humanizations, that represents trajectories in the sequence space, leading to high-scoring sequences. The proposed SD/SAMC methods for optimizing the score produce a huge number of sequences beyond the threshold score, that could be considered as candidate humanized sequences. Most often, the SAMC procedure finds better solutions than the (much faster) SD, that might get trapped in suboptimal sequences. However, considering that in a humanization protocol it is more interesting to get good sequences with a minimal amount of mutations of the original sequence, rather than a global optimum, we have focused mainly in the SD trajectories originating at the different starting murine sequences. In all cases that we have analyzed, there are several common mutations between the last (and best) sequence in the SD procedure and the experimentally humanized sequence; the latter is not visited along the SD trajectory, even if this is not necessarily something to worry about, since there is no reason for the experimentally humanized sequence to be the only acceptable one. Interestingly, the optimal sequence that we find by SD is generally recognized by protein-BLAST as "more human" than the corresponding experimentally humanized one; this suggests that intermediate sequences found in the SD trajectory, with lower humanness score, could represent suitable candidates as well, and their higher similarity to the original functional sequence should reduce the risk of loosing stability or solubility.

So, within a humanization protocol, the advantage of our approach over CDR-grafting is that it proposes a precise set of candidate sequences, at increasing distance from the murine one and increasing humanness score, instead of requiring the introduction of arbitrary, manual back-mutations; such sequences can be shortlisted by further modelling of their structure and stability, while the final choice relies on the experimental result from the best hits. We stress that the method we propose represents an alternative approach to the usual grafting on germline sequences, since it deals with the statistical properties of the ensemble of mature sequences, and its performance can only improve as more sequences are collected, yielding better statistics. Indeed, the ensemble of rearranged human sequences that we use as learning set presents some of the desired properties for therapeutic antibodies, due to their nature (we expect that they share at least solubility and low immunogenicity); thus, as far as these properties can be recovered by a statistical model based on residue-residue interaction, we expect that an increase of the number of sequences in the database will yield high scoring sequences being indeed better humanized sequences.

Moreover, inferring a score from a database of experimentally verified sequences represents a flexible approach, that can be adapted to different learning databases. In perspective, this involves precision-medicine applications, resorting to individual antibody repertoires, that are becoming available through Rep-Seq techniques. Also, it can be easily adapted to veterinary drugs development, by changing the human learning dataset to the appropriate animal one. With the growing wealth of sequence databases, statistical-inference methods could become an increasingly relevant tool, with a range of applications that is still to be explored.

Methods

We report here a general description of the methods: further details can be found in SI.

Database preparation.

Learning databases. We download from the IMGT/LIGM-DB server two databases with the whole set of human, rearranged, cDNA, VH sequences (11463 units) and VL sequences (5546 units), respectively, in the IMGT format. We extract and annotate each sequence with a unique identification strings that combines several fields of the IMGT record. To align the sequences, we resort to the ANARCI tool[18] (version 1.1) with the AHo numbering scheme[19], that is structurally motivated and basically free from insertions. We build a combined VH-VL database by matching VH and VL sequence according to their identification string, and joining them in a unique, aligned sequence of fixed length L = 298 (including gaps: the AHo scheme aligns both VH and VL in frameworks of 149 positions). All cases where the matching is not unique are removed. This yields a database of 1309 joint VH-VL sequences.

We perform the same steps on two databases with the whole set of mouse ("mus musculus"), rearranged, cDNA, VH (8389 units) and VL sequences (1514 units), both downloaded from the IMGT/LIGM-DB server, ending up with a combined database of 373 aligned sequences. Notice that we use the mouse learning database only for the classification with two reference distributions, while the rest of our results are based on just the human learning dataset.

Test databases. We download from the DIGIT server[20] the whole database of matching human VH-VL sequences (3322 sequences), aligned according to the Kabat scheme. We remove the alignment, split the sequences into separate VH and VL chains, and filter on their length as above. We perform the same steps for the database of murine VH-VL matching sequences (1933 units.) Then, we use ANARCI to align the human VH and VL files according to the AHo scheme, and eliminate repeated sequences both within the DIGIT VH or VL files, as well as between these and the corresponding murine DIGIT aligned files, and the aligned human and mouse learning databases. Finally, we combine the VH and VL sequences, obtaining a database of 1388 sequences. The same procedure is repeated for the murine VH and VL files, yielding a combined database of 1379 sequences.

Humanized and Therapeutic antibodies database. We extract from DrugBank a list of therapeutic antibodies whose sequences are publicly available, reported in SI Table S1. For these sequences, we have also searched publicly available information on immunogenicity (as measured as the frequency of appearance of antidrug antibodies)[21–23]. The complete list of sequences, together with their reported immunogenicity, can be found in SI file "Therapeutic_Ab.txt". Also, we compiled from literature[24–30] a list of pairs of corresponding murine and experimentally humanized sequences, reported in SI Table S2.

Definition of the CDRs. We define the CDR regions according to the IMGT scheme: in the AHo layout, this imply the following definition: for the VH region (residues 1–149): 27–40 (CDR1), 58–68 (CDR2), 107–138 (CDR3); for the VL region (residues 150–298): 176–189 (CDR1), 207–217 (CDR2), 257–287 (CDR3). These positions are calculated with the following protocol: rather than resorting to the correspondence table reported in www.bioc.uzh.ch/plueckthun/antibody/Numbering/NumFrame.html we align some input sequences in the AHo scheme, and then query the IMGT-VQUEST[31] server, to identify the CDR regions, and find their position in the alignment.

Sequence Classifiers.

We compare two different approaches, one based on the identity between pairs of sequences, and the latter based on the inference of a statistical model describing each (human or murine) sequence distribution.

Classifiers based on sequence identity. We define the distance between two sequences $\mathbf{A}^1 = \{A_i^1, i = 1, \ldots, L\}$, $\mathbf{A}^2 = \{A_i^2, i = 1, \ldots, L\}$, where L is the length of the alignment (including gaps) in the AHo numbering scheme, as: $d_{\mathbf{A}^1,\mathbf{A}^2} = \sum_{i=1}^{L}(1 - \delta_{A_i^1 A_i^2})$, where A_i^m indicates the amino acid type at position i of sequence m and $\delta_{X,Y}$ is the Kronecker delta, equal to one or zero according to whether the residues X and Y are identical or not. From this, several different proximity scores can be defined, of the kind: $T_k(\mathbf{A}) = \frac{1}{k}\sum_{\mathbf{a}\in\mathcal{N}_k^{(h)}(\mathbf{A})} d(\mathbf{A}, \mathbf{a})$ where $\mathcal{N}_k^{(h)}$ is the set of k sequences in the human learning dataset that are most similar to the query sequence $\mathbf{A}$. Hence, $T_1(\mathbf{A})$ is the distance of $\mathbf{A}$ from the closest sequence in the human learning database, while we call T_{all} the average distance of $\mathbf{A}$ from the learning ensemble of human sequences: $T_{all} \equiv \overline{d}^{(h)}(\mathbf{A}) = T_{M_h}(\mathbf{A})$, with M_h the number of sequences in the human learning database. Analogously, it is possible to define the equivalent quantities (e.g. $\overline{d}^{(m)}(\mathbf{A})$) referred to the the murine learning ensemble.

The classification with one reference distribution uses $T_k(\mathbf{A})$ as the score of $\mathbf{A}$. A threshold score separating the human from non-human (mouse) class is defined for each method, by optimizing its performance on the test databases: each value of the threshold involves a different number of False Positives (*FP*) and True Positives (*TP*), which correspond to a point in the *FPR-TPR* plane, where the False Positive Rate (*FPR*) and True Positive Rate (*TPR*) are defined as: $TPR = TP/P$; $FPR = FP/N$, where P and N are, respectively, the number of human sequences (considered as Positive events), and of murine sequences (considered as Negatives) in the test databases. Varying the values of the threshold score, a curve (ROC curve) is drawn in the *FPR-TPR* plane, characterizing the goodness of the classifier: the bigger the area under the curve, the better is the classifier. The best threshold value can be determined as the one yielding the point on the curve that maximizes the Youden's index (that in general, is also very close to the one maximizing the Matthews Correlation Coefficient): $Y = TP/(TP + FN) + TN/(TN + FN) - 1$ or the Matthews Correlation Coefficient $M = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$. Notice that the threshold score identified in this way is maintained when analyzing the therapeutic antibodies. The classification with two reference distributions does not need the definition of a threshold score: it is based on the difference: $s(\mathbf{A}) = \overline{d}^{(h)}(\mathbf{A}) - \overline{d}^{(m)}(\mathbf{A})$, with positive values indicating that the sequence $\mathbf{A}$ is closer to the human ensemble (and therefore, human) while negative values correspond to murine sequences.

Classifiers based on inference of a probabilistic model. We follow Baldassi *et al.*[10] and infer a multivariate Gaussian distribution from each VH, VL and combined VHVL learning databases, using uninformative prior distributions. From this, we calculate the posterior predictive distribution, that results to be a multivariate Student distribution, and use it to score the sequences in the test datasets. We sketch below the main tenets of the approach; more details can be found in SI Methods.

We start by mapping the L-residues-long, aligned sequences of the database (made up by M sequences, and drawn from a $Q = 20$ letters alphabet) to a binary sequence of $N = QL$ bits $\{x_i = 0, 1, i = 1, \ldots, N\}$, that, in block of Q bits, represent all the amino acids. As in ref.[10] we assume that each of the M sequences in the database is drawn from a normal distribution with parameters μ, Σ (thus promoting x_i^m to be real numbers):

$$p(x^m|\mu, \Sigma) = \mathcal{N}(\mu, \Sigma) \tag{1}$$

and we assume a Normal Inverse Wishart prior distribution for the parameters μ, Σ: $p^{pr}(\mu, \Sigma) = \mathcal{NIW}(\eta, \kappa, \Lambda, \nu) = \mathcal{N}\left(\mu|\eta, \frac{\Sigma}{\kappa}\right)\mathcal{IW}(\Sigma|\Lambda, \nu)$. Using Bayes theorem, the posterior distribution for μ, Σ, given the data X can be calculated, yielding again a NIW distribution with new parameters η', κ', Λ', ν':

$$p^{post}(\mu, \Sigma|X) \propto p(X|\mu, \Sigma)p^{pr}(\mu, \Sigma) = \mathcal{NIW}(\eta', \kappa', \Lambda', \nu') \tag{2}$$

where the expression of η', κ', Λ', ν' can be found in Eq. (10) of ref.[10] as a function of the empirical average $\overline{x}$ and empirical covariance $\overline{C}_{i,j}$ of the data. In a Bayesian approach, we derive the posterior predictive distribution for each new sequence $y = \{y_i, i = 1, \ldots, N\}$, by integrating on μ, Σ the joint probability: $p(y, \mu, \Sigma|X) = p(y|\mu, \Sigma)p^{post}(\mu, \Sigma|X)$. Plugging Eqs (1) and (2) into the above equation, we get

$$p(y, \mu, \Sigma|X) = \rho(y)\ \mathcal{NIW}\ (\mu, \Sigma|\eta'', \kappa'', \Lambda'', \nu'') \tag{3}$$

where the expressions for η'', κ'', Λ'', ν'' and ρ are reported in SI Methods. Upon integrating the above equation on μ, Σ, we get the posterior predictive distribution of a new sequence y, given the database of sequences X, as the multivariate t-distribution probability density:

$$p(y|X) = t_N\left(\frac{M}{1 - \lambda} + 2, \langle\mu\rangle_{post}, \left(1 + \frac{1 - \lambda}{M}\right)\langle\Sigma\rangle_{post}\right) \tag{4}$$

with: $\langle\mu\rangle_{post} = \lambda\eta + (1 - \lambda)\overline{x}, \langle\Sigma\rangle_{post} = \lambda U + (1 - \lambda)\overline{C} + \lambda(1 - \lambda)\ (\overline{x} - \eta)\ (\overline{x} - \eta)^T$. We choose η and U as those corresponding to the mean and covariance estimates of a uniformly distributed sample.

We use the logarithm of $p(y|X)$ in Eq. (4) as a score of the humanness of any given sequence y, and call it the "MG score". Notice that $p(y|X)$ is a probability density, and not a probability: as such, it is not bound between 0 and 1, and actually, due to its strong localization in the high dimensional sequence space, it will greatly exceed 1.

Finally, we have to choose a value for λ to be used in our inference, to set the best amount of regularization λU that should be added to the empirical covariance to optimize the statistical model. We do so by analyzing the different ROC curves, obtained at different values of λ in classifying the test databases (see the previous section for the definition of the ROC curve), and choosing the value of λ yielding the curve with the maximal area under it. Finally, for the classification with one reference distribution, we select, as the threshold score, the one corresponding to the point, on the ROC curve for the test database, with the highest value of the Youden's coefficient. Notice that the threshold score identified in this way is maintained when analyzing the therapeutic antibodies, and in the humanization protocol.

When classifying with two distributions, we fix the λ for both the murine and human statistical models as explained above. Then, we simply score each query sequence in the test databases with both statistical models, classifying it as human or murine depending on the which of the two scores is higher.

Classification without correlations is performed by maintaining the same value of Ω and neglecting correlations between blocks of binary variables representing residues at different positions along the sequence; this ensures that Σ^{-1} will be block-diagonal as well, with no interactions between residues. We optimize λ as before, obtaining $\lambda = 0.027$. Analogously, classification without correlations between VH and VL regions is performed by maintaining the same value of Ω and asking that $\Sigma ij = 0$ if k, l belong to the VH and VL region, respectively. Again, the resulting Σ^{-1} will be block-diagonal, with no interactions between residues belonging to different variable regions. In this case, $\lambda = 0.067$.

Humanization of murine sequences. We use the negative MG score, $E = -\log(p(y|X))$, see Eq. 4, as the objective function (the "Energy", in analogy with statistical physics), that thus has to be minimized. We start from a murine sequence and, keeping fixed the positions corresponding to the CDR regions of the antibody, to ensure that it will continue recognizing the antigen, we mutate the other residues to make the sequence more human like. To this end, we perform two different strategies: 1) A steepest descent approach (SD), consisting in choosing, at each step, the mutation, at any position, that causes the biggest decrease in E. The process stops when all possible mutations at all sites increase the "energy"; 2) A Simulated Annealing Monte Carlo approach (SAMC), where we propose a random mutation at a random position, and accept it according to the Metropolis scheme. The simulated annealing starts at a high temperature, $T = 25.1$, and lowers the temperature in $\Delta T = 0.5$, every $3.74 \cdot 10^6$ MC steps, until reaching $T = 0.1$. Here the temperature is just a parameter that controls the ability of the system to jump away from local minima and freely move through the sequence space (at high T), or to get trapped and explore the basins surrounding a minima, at low T.

The first technique finds the closest local minimum of the objective function in the neighborhood of the starting sequence, and the optimal path to reach it. Simulated Annealing, starting from high temperatures, explores wider regions of the sequence space, and in general finds the deepest minima, corresponding to higher humanness score.

We have tried both methods to humanize the 7 different murine sequences contained in the murine-humanized pairs dataset, SI Table S2. Since we know the corresponding experimentally humanized partner for each one, we can even control to what extent our theoretical humanized sequence matches its experimental counterpart.

Data Availability

The learning and test datasets, that we have compiled from public sources and analysed during the current study, are available from the corresponding author on reasonable request.

References

1. Williams, D. G., Matthews, D. J. & Jones, T. Humanising Antibodies by CDR Grafting. In *Antibody Engineering*, 319–339 (Springer, https://doi.org/10.1007/978-3-642-01144-3_21 2010).
2. Abhinandan, K. & Martin, A. C. Analyzing the "Degree of Humanness" of Antibody Sequences. *J. Mol. Biol.* **369**, 852–862, https://doi.org/10.1016/j.jmb.2007.02.100 (2007).
3. Gao, S. H., Huang, K., Tu, H. & Adler, A. S. Monoclonal antibody humanness score and its applications. *BMC Biotechnology* **13**, 1–12, https://doi.org/10.1186/1472-6750-13-55 (2013).
4. Lazar, G. A., Desjarlais, J. R., Jacinto, J., Karki, S. & Hammond, P. W. A molecular immunology approach to antibody humanization and functional optimization. *Mol. immunology* **44**, 1986–1998 (2007).
5. Choi, Y., Hua, C., Sentman, C. L., Ackerman, M. E. & Bailey-Kellogg, C. Antibody humanization by structure-based computational protein design. *mAbs* **7**, 1045–1057, https://doi.org/10.1080/19420862.2015.1076600 (2015).
6. Seeliger, D. Development of scoring functions for antibody sequence assessment and optimization. *PloS one* **8**, e76909 (2013).
7. De Groot, A. S. & Martin, W. Reducing risk, improving outcomes: bioengineering less immunogenic protein therapeutics. *Clin. Immunol.* **131**, 189–201 (2009).
8. Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody-antigen recognition. *Front. immunology* **4** (2013).
9. Morcos, F. *et al*. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Nat. Acad. Sci.* **108**, E1293–E1301 (2011).
10. Baldassi, C. *et al*. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PloS one* **9**, e92721 (2014).
11. Asti, L., Uguzzoni, G., Marcatili, P. & Pagnani, A. Maximum-entropy models of sequenced immune repertoires predict antigen-antibody affinity. *PLoS Comput. Biol.* **12**, e1004870 (2016).
12. Eddy, S. R. Profile hidden Markov models. *Bioinforma.* **14**, 755–763, https://doi.org/10.1093/bioinformatics/14.9.755 (1998).
13. Wishart, D. S. *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **34**, D668–D672 (2006).
14. Law, V. *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **42**, D1091–D1097 (2013).
15. Wishart, D. S. *et al*. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research* **36**, D901–D906 (2007).
16. Wang, P. *et al*. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS computational biology* **4**, e1000048 (2008).
17. Wang, P. *et al*. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC bioinformatics* **11**, 568 (2010).
18. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinforma.* **32**, 298–300, https://doi.org/10.1093/bioinformatics/btv552 (2016).
19. Honegger, A. & Plückthun, A. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *J. Mol. Biol.* **309**, 657–670, https://doi.org/10.1006/jmbi.2001.4662 (2001).
20. Chailyan, A., Tramontano, A. & Marcatili, P. A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res.* **40**, D1230–D1234, https://doi.org/10.1093/nar/gkr806 (2011).
21. Hwang, W. Y. K. & Foote, J. Immunogenicity of engineered antibodies. *Methods* **36**, 3–10 (2005).
22. Harding, F. A., Stickler, M. M., Razo, J. & DuBridge, R. The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the CDR regions. *mAbs* **2**, 256–265 (2010).

23. Baker, M., Reynolds, H. M., Lumicisi, B. & Bryson, C. J. Immunogenicity of protein therapeutics: the key causes, consequences and challenges. *Self/nonself* **1**, 314–322 (2010).
24. Gabbard, J. *et al.* A humanized anti-M2 scFv shows protective *in vitro* activity against influenza. *Protein Eng. Desi. Sel.* **22**, 189–198, https://doi.org/10.1093/protein/gzn070 (2008).
25. Hu, W.-G., Yin, J., Chau, D., Negrych, L. M. & Cherwonogrodzky, J. W. Humanization and Characterization of an Anti-Ricin Neutralization Monoclonal Antibody. *PLoS ONE* **7**, e45595, https://doi.org/10.1371/journal.pone.0045595 (2012).
26. Asano, R. *et al*. Construction and humanization of a functional bispecific EGFRxCD16 diabody using a refolding system. *FEBS J.* **279**, 223–233, https://doi.org/10.1111/j.1742-4658.2011.08417.x (2012).
27. Jia, X. *et al.* A humanized anti-DLL4 antibody promotes dysfunctional angiogenesis and inhibits breast tumor growth. *Sci. Reports* **6**, https://doi.org/10.1038/srep27985 (2016).
28. Shaw, S. *et al.* Discovery and characterization of olokizumab. *mAbs* **6**, 773–781, https://doi.org/10.4161/mabs.28612 (2014).
29. Margreitter, C., Mayrhofer, P., Kunert, R. & Oostenbrink, C. Antibody humanization by molecular dynamics simulations-in-silicoguided selection of critical backmutations. *J. Mol. Recognit.* **29**, 266–275, https://doi.org/10.1002/jmr.2527 (2016).
30. Shembekar, N. *et al.* Humanized antibody neutralizing 2009 pandemic H1N1 virus. *Biotechnol J.* **9**, 1594–1603, https://doi.org/10.1002/biot.201400083 (2014).
31. Brochet, X., Lefranc, M.-P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized VJ and VDJ sequence analysis. *Nucleic acids research* **36**, W503–W508 (2008).

Acknowledgements

The authors acknowledge MINECO grant FIS2014-55867-P (P. B.), FIS2017-87519-P (P. B.) and FIS2015-65078-C2-2-P (S. P. G.), and financial support by the Aragon Regional Government through Research Group E24/3 (P. B., S. P. G.). T. D. M. was partially supported by a Diatheva grant.

Author Contributions

P.B. and M.M. designed the approach, P.B. performed the analytical calculations, T.D.M., M.M., P.B. identified and processed the datasets, A.C.A. and S.P.G. implemented the code and run the simulations, all the authors analyzed and discussed the results, P.B., M.M. wrote the article. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-32986-y.

Competing Interests: M.M. is the scientific director of Diatheva srl, and holds shares in the company; T.D.M. was partially supported by a Diatheva grant. The rest of authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.