

OBTAINING AN IMPROVED THERAPEUTIC LIGAND

The present invention relates to obtaining an improved therapeutic ligand, in particular by determining how an existing or candidate ligand can be modified to improve binding of the ligand at a binding site on a target protein or by aiding the *de novo* design of a candidate ligand as a precursor to a therapeutic.

Therapeutic molecules (ligands) fall into two distinct classes: chemical entities (or novel chemical entities, NCEs) and biologicals. The former are low molecular weight organic compounds, typically of molecular weight of 500 Daltons or less, that have been chemically synthesized or isolated from natural products. These are typically derived from starting chemicals or 'hits' that are discovered by screening chemical or natural product libraries. Such hits typically have sub-optimal binding affinity for the target and considerable trial and error in chemical modification is required in order to obtain better affinity for the target (typically of affinity constant (K_D) low micromolar or less). It is preferable that the hit has a lower molecular weight, say 300 Daltons or less, so that subsequent chemical modification does not exceed the 500 Dalton limit. These hits are often referred to as 'fragments'. Optimisation of the hit to obtain a candidate therapeutic or lead molecule is greatly enhanced by structural information; for instance by obtaining an x-ray crystallographic structure of the protein in co-complex with the hit molecule or fragment. Such data provides insight into where on the target protein the small molecule binds and importantly indicates how atomic interactions between the two account for binding. Furthermore the topographical nature of the protein surface immediately surrounding the bound hit is revealed; and particularly if it is a cleft or a pocket, the structure will suggest how the hit might be elaborated to better fill the space within the pocket and how to make further interactions with the protein and hence improve binding affinity and specificity.

There are a number of computer based algorithms available to assist the medicinal chemist in making rational choices for chemical elaboration of the hit. These are either physics based methods that attempt to calculate the free energy of binding between the small molecule and protein from first principles (e.g. Schrodinger (RTM) suite of software) or are statistical potential methods that rely on a database of atomic interactions extracted from collections of protein – small molecule structures (e.g. SuperStar).

Biologicals are large peptide or protein molecules (of molecular weight greater than

1000 Daltons). They are often antibodies or antibody like molecules that recognize and bind to a target molecule, usually with better affinity and specificity compared to NCEs (K_D low nanomolar or less). They may also be other types of protein molecules such as hormones, cytokines, growth factors or soluble receptors.

5 The binding of a candidate biological therapeutic molecule to a binding site can be modified by mutating the candidate therapeutic molecule. This may be required to improve the binding affinity or alter the binding specificity. However, it is relatively time-consuming to perform the mutation and to test the binding efficiency of the mutated molecule. Many different mutations may be required before improvements in
10 binding efficiency are obtained.

It is known to use computers to predict what kind of modifications might be most effective. However, a given molecule can be modified in a vast number of ways and it is difficult to configure a computer so that the prediction can be achieved reliably in a practical period of time.

15 Laskowski R A, Thornton J M, Humblet C & Singh J (1996) "X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins", Journal of Molecular Biology, 259, 175-201 discloses a computer-based method for identifying favourable interaction regions for different atom types at the surface of a protein, such as at a dimer interface or at a
20 molecular-recognition or binding site. The Laskowski et al predictions are based on a database of empirical data about non-bonding intra-molecular contacts observed in high-resolution protein structures.

In the approach of Laskowski et al, the 20 amino acids are broken up to yield a total of 488 possible 3-atom fragments. Taking chemical similarities into account these are
25 reduced to a set of 163 fragment types that sub-divide the database. Each fragment contains a first atom (referred to as "position 3") with two further atoms defining triangulation (or spatial normalization) positions. A density function is derived by recording the various positions at which an atom (which may be referred to as a "second atom") is found to be in a non-bonding intra-molecular contact with the first atom of a
30 3-atom fragment.

A predicted favourable interaction region for a given atom type is obtained in Laskowski et al by transplanting density functions into the binding site. Each density function is transplanted such that the coordinates of the three atoms of the 3-atom fragment corresponding to the density function are superimposed on the coordinates of

a corresponding 3-atom fragment in the binding site. Where density functions from different 3-atom fragments in the binding site overlap, an average “density” is used to predict the favourable interaction region.

5 The approach of Laskowski et al is relatively complex and discards potentially useful data. The density functions of Laskowski et al are obtained by populating a 3-D grid with the positions of second atom contacts for each fragment type. A different grid is used for each of the 163 fragment types. Data for each second atom type is then mathematically transformed to give the density function. Using the fragment definitions of Laskowski et al, fragment type can be shared by different atoms on the same residue and by atoms on different residues. When the Laskowski et al database has been built on 10 these fragment types there is an over-abundance of main chain fragments that requires a down-weighting at the stage of transplanting the density functions into the binding site. Furthermore, a given fragment type can include several actual fragments with subtle differences in bond lengths and angles and concomitant differences in second atom 15 distributions which become masked when combined in these divisions.

Short range secondary structure in the proteins used for deriving the empirical data in Laskowski et al can lead to bias and reduces the efficiency with which the empirical data indicates favourable interaction regions.

20 It is an object of the invention to address at least one of the problems with the art discussed above.

According to an aspect of the invention, there is provided a method for designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target, comprising:

- 25 a) identifying a target list of atoms forming the surface of the target binding site;
b) identifying each atom, hereinafter referred to as a theta atom, in the target list, as a particular theta atom type;
c) extracting from a structural database of biological macromolecules, information about non-bonding, intra-molecular or inter-molecular atom to atom contacts, where 30 the first atom in a contacting pair of atoms is of a particular theta atom type and the opposing, second atom of the pair, hereinafter referred to as an iota atom, is of a particular iota atom type, said information comprising spatial and/or contextual data about the iota atom relative to the theta atom, and said data collected for a plurality of contacts of the given theta atom type from the said database is hereinafter referred to as

a theta contact set;

d) for each theta atom identified in the target list in step b), superimposing in or around the target binding site data relating to a given iota atom type, or a predetermined group of related iota atom types, from the corresponding theta contact set extracted in step c);

e) combining and / or parsing the superimposed data in such a way as to predict one or more favoured regions of the binding site where the given iota atom type, or the predetermined group of related iota atom types, has high theoretical propensity; and

f) with a candidate ligand notionally docked into the binding site, comparing the type and position of one or more of the atoms of the candidate ligand with the predicted favoured regions for the respective iota atom types, to identify a modification to the candidate ligand, in terms of alternate and/or additional candidate ligand atoms, that will produce a greater intersection between the alternate and/or additional candidate ligand atoms and the respective iota atom type favoured regions, leading to an improvement in the affinity of the modified candidate ligand to the binding site compared to the unmodified candidate ligand;

wherein each non-bonding intra-molecular or inter-molecular contact in the database is defined as a contact between opposing residues of a protein fold or between opposing monomer units of a macromolecular fold or between two interacting macromolecular partners and is specifically between a theta atom on one side of the fold or first interacting partner and an iota atom on the opposing side or second interacting partner; in an instance where the following condition is satisfied:

$s - R_w \leq t$, where s is the separation between the two atoms of the contact, R_w is the sum of the van de Waals radii of the two atoms of the contact, and t is a predetermined threshold distance; and

wherein the theta atom type is identified uniquely in step b) such that there is no intersection between the data of a theta contact set extracted in step c) for a given theta atom type and the data of any other theta contact set extracted in step c) for any other theta atom type, apart from data concerning contacts involving the given theta atom as the iota atom.

Thus, each target atom type in the binding site is classified uniquely and is associated with information about a set of contacts extracted from the structural database that is unique and which does not overlap with the set of contacts associated with any other atom type (apart from those contacts which involve the target atom type itself as the iota

atom). This means that a distribution of theoretical locations for a given iota atom type, or a predetermined group of related iota atom types, determined based on one target atom in the binding site may be combined (e.g. by summing) more efficiently (e.g. without weighting) with a distribution of theoretical locations for an iota atom type, or
5 predetermined group of related iota atom types, determined based on another target atom in the binding site, for example to provide an improved prediction of one or more favoured regions for the iota atom type or predetermined group of related iota atom types. Also because each target atom in the binding site is classified uniquely, there are no variations in bond lengths or angles to consider and hence the theoretical location of
10 a given iota atom is more precise.

In an embodiment, simple rules are applied to uniquely identify the neighbouring atoms for the purposes of triangulation. No assumptions need to be made about the chemical nature of neighbouring atoms, which is necessary for example where contact types are characterized in terms of the 163 3-atom fragment types of Laskowski et al.

15 In an embodiment, the spatial data extracted in step c) defines the position of each iota atom specified in the theta contact set by geometrical reference to the position of the theta atom and to the positions of third and fourth atoms, wherein the third atom is covalently bonded to the theta atom and the fourth atom is covalently bonded to the third atom. In an example of such an embodiment, for each iota atom specified in the theta contact set, said spatial data extracted in step c) further defines the position of fifth
20 and sixth atoms by geometrical reference to the position of the theta atom and to the positions of the third and fourth atoms, wherein the fifth atom is covalently bonded to the iota atom and the sixth atom is covalently bonded to either the fifth atom or the iota atom.

25 In an embodiment, the superimposition in or around the target site of step (d) comprises: parsing the theta contact set to extract spatial data for contacts comprising the given iota atom type or one or more of the predetermined group of related iota atom types; and plotting this spatial data to determine theoretical locations representing where each iota atom type, or each of the one or more of the predetermined group of related
30 related iota atom types, would be located if: i) the theta atom of the contact were located at the position of the corresponding theta atom in the target binding site; and ii) the third and fourth atoms of the contact were located at the positions of the third and fourth atoms of the corresponding theta atom in the target binding site. In an embodiment, the spatial data is parsed against the contextual data before the plotting step.

In an embodiment, a region in which a density of theoretical locations for the iota atom type (or one or more of the predetermined groups of related iota atom types) is above a predetermined threshold is identified as one of the favoured regions. In an example of such an embodiment, theoretical locations for the given iota atom type, or for one or more of the predetermined group of related iota atom types, are determined for a plurality of theta atoms on the target list and a region in which a density of the cumulative theoretical locations is above the predetermined threshold is identified as one of the favoured regions.

Thus, theoretical locations are combined cumulatively from different atoms in the binding site before the density of theoretical locations is obtained for the purposes of predicting favoured regions. This results in a more accurate statistical representation of the probability of a given iota atom type, or in a given group of related iota atom types, being positioned at a given location because it takes into account the contributions from all relevant atom types in the binding site in a proportionate and unbiased manner. In Laskowski et al., in contrast, the density functions are derived for groups of 3-atom fragments. Each atom may be associated with several different groups of 3-atom fragments and so it is not possible simply to add together density functions in a manner comparable with embodiments of the present invention. Instead, it is necessary to perform weighting and/or averaging before combining density functions, which increases complexity and/or reduces accuracy.

In an embodiment only contacts between atoms that are separated from each other by four residues or more are used for identifying favoured regions. This significantly reduces or avoids bias due to short range secondary structure. In an embodiment, the contact data predominantly represents long-range, across-fold protein data.

According to a further aspect of the invention, there is provided a method of generating a database for use in a method for designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target, comprising:

analysing the relative positions of atoms in each of a plurality of proteins or other biological macromolecules in order to identify instances of a non-bonding intra-molecular contact between a first atom, referred to as a theta atom, and a second atom, referred to as an iota atom, of the protein or macromolecule; and

generating a database that for each identified contact specifies: the type of the theta

atom, the type of the iota atom, and the position of the iota atom relative to the theta atom;

wherein a non-bonding intra-molecular contact is defined as an instance where the following conditions are satisfied:

5 $s - R_w \leq t$, where s is the separation between the theta and iota atoms, R_w is the sum of the van de Waals radii of the theta and iota atoms, and t is a predetermined threshold distance of typically 2.5 angstroms and preferably 0.8 angstroms; and

wherein in the case of proteins, the theta and iota atoms are on amino acid residues separated from each other by at least four residues on a linear polypeptide or are on
10 separate polypeptide chains.

According to a further aspect of the invention, there is provided a method of generating a database for use in a method for designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target,
15 comprising:

analysing the relative positions of atoms in each of a plurality of proteins or other biological macromolecules in order to identify instances of a non-bonding intra-molecular contact between a first atom referred to as a theta atom, and a second atom, referred to as an iota atom, of the protein or macromolecule; and

20 generating a database that for each identified contact specifies: the type of the theta atom, the type of the iota atom, and the position of the iota atom relative to the theta atom;

wherein a non-bonding intra-molecular contact is defined as an instance where the following condition is satisfied:

25 $s - R_w \leq t$, where s is the separation between the theta and iota atoms, R_w is the sum of the van de Waals radii of the theta and iota atoms, and t is a predetermined threshold distance of typically 2.5 angstroms and preferably 0.8 angstroms; and

wherein the method comprises sub-dividing the database to form groups of identified contacts in which the theta atom is one and only one of the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins and the iota atom is in one and only
30 one of a plurality of non-overlapping groups obtained by sorting the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins into groups based on chemical similarity.

Embodiments of the invention will now be described, by way of example only, with

reference to the accompanying drawings in which corresponding reference symbols represent corresponding parts, and in which:

Figure 1 is a schematic illustration of an example nomenclature for atoms at a non-bonding intra-molecular or inter-molecular contact and neighbouring atoms that are used for coordinate normalization;

Figures 2-6 illustrate a process of coordinate normalization for an atom in a non-bonding intra-molecular or inter-molecular contact;

Figure 7 is a flow chart illustrating steps in a method of designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target;

Figure 8 is a computer generated visualization depicting a light chain threonine 30 to arginine 30 mutation;

Figure 9 is a computer generated visualization depicting a light chain arginine 54 to serine 54 mutation;

Figure 10 is a computer generated visualization depicting a light chain serine 56 to isoleucine 56 mutation;

Figure 11 is a computer generated visualization depicting a light chain serine 60 to aspartate 60 mutation;

Figure 12 is a computer generated visualization depicting a light chain threonine 72 to arginine 72 mutation;

Figure 13 is a computer generated visualization depicting the combination of 5 mutations in antibody 496i light chain resulting in a 180-fold improved affinity to IL17F;

Figure 14 is a flow chart illustrating steps in a method of predicting the effects of point mutations at the VH-VL interface of a Fab;

Figure 15 is a computer generated visualisation depicting a heavy chain threonine 71 to arginine 71 mutation;

Figure 16 is a computer generated visualisation depicting a light chain serine 107 to glutamic acid 107 mutation;

Figure 17 is a computer generated visualisation depicting a light chain threonine 109 to isoleucine 109 mutation;

Figure 18 is a computer generated visualisation depicting the combination of three mutations in Fab X resulting in a T_m of 81.2°C.

The Worldwide Protein Data Bank (wwPDB) maintains an archive of macromolecular structural data that is freely and publicly available to the global community. By May 2013 this dataset had reached the milestone of 90 000 structures. Most of these macromolecules are proteins of which the majority have been determined by X-ray crystallography. Deposited data thus contains three dimensional data at the atomic level in the form of Cartesian coordinates of individual atoms that make up the respective protein structure.

The inventors hypothesised that it is possible to extract useful information from this archive which could be applied to aid the design of novel therapeutics. The polypeptide chains of a nascent protein fold into complex three dimensional tertiary and quaternary structures in a remarkably reproducible manner to yield the mature protein. Interactions affecting the formation of the secondary structure of proteins, elements such as helices, beta-sheets and turns, are known. However, rules predicting the higher orders of protein folding are poorly understood. Nonetheless, the inventors have realised that there must be precise rules that govern the interaction of non-bonding but “contacting” atoms, either within the same molecular, for example on opposing faces of a protein fold, or on different molecules.

In an embodiment, a structural database of biological macromolecules (e.g. the wwPDB) is analysed to extract such rules, and the rules are applied to facilitate drug discovery. An example of such a process is described below.

In an embodiment, non-bonding pairs of contact atoms (referred to respectively as “theta” and “iota” atoms) are identified for each macromolecule (e.g. protein), or a subset of fewer than all of the macromolecules, in the structural database of macromolecules (e.g. the wwPDB). Such contacts may occur for example between opposing residues of a protein fold or between opposing monomer units of a macromolecular fold (between separate chains of a macromolecular structure) or between two interacting macromolecular partners. Each contact is classified as being between a theta atom on one side of the fold or first interacting partner and an iota atom on the opposing side or second interacting partner.

In an embodiment, the non-bonding intra-molecular or inter-molecular contacts are defined as an instance where the following condition is satisfied: 1) $s - R_w \leq t$, where s is the separation between the two atoms of the contact, R_w is the sum of the van de Waals radii of the two atoms of the contact, and t is a predetermined threshold distance; and, optionally, the following condition also: 2) the two atoms of the contact are

separated from each other by at least four residues along a linear polypeptide chain or are on separate polypeptide chains. In an embodiment, the predetermined distance is 2.5 angstroms. In another embodiment, the predetermined distance is 1.5 angstroms. In another embodiment, the predetermined distance is 1.0 angstroms. In another
5 embodiment, the predetermined distance is 0.8 angstroms.

In the description below, any reference to “contact” is understood to mean “non-bonding intra-molecular contact or inter-molecular contact” according to the definition given above.

Databases such as the wwPDB may have information about proteins that are very
10 similar to each other and/or which have related structures. In an embodiment, the database is parsed in order to avoid/reduce bias caused by such similarities/relationships. In an embodiment, the parsing is performed based on primary sequence homology, for example such that only one representative structure of each family of similar/related proteins is selected for analysis. Additionally or alternatively,
15 one or more further selection criteria may be used, for example high resolution and low temperature factor structures may be incorporated.

In an embodiment, a secondary database is constructed starting from the (primary) structural database of biological macromolecules (e.g. the wwPDB). The secondary database comprises information about the non-bonding intra-molecular or inter-
20 molecular contacts. In an embodiment, the secondary database comprises information about more than 1 million contact pairs, optionally more than 5 million contact pairs, optionally more than 11 million contact pairs. In one embodiment, the secondary database comprises information from more than 15 million contact atom pairs, extracted from around 20 000 non-homologous proteins.

25 In an embodiment, the secondary database contains information about the precise atom types of the contact pair. In an embodiment, the secondary database contains spatial data defining the three dimensional relationship of the theta atom to the iota atom. In an embodiment, the secondary database also contains contextual data concerning the local environment of the contact. In an embodiment, the contextual data
30 contains information concerning the local environment of each contact pair, including one or more of the following in any combination: secondary structure, amino acid types or other monomer types comprising the contact pair, adjacent monomer units and/or local geometry thereof in a polymer chain either side of the contact, adjacent amino acids in a polypeptide chain on either side of the contact, local geometry of the said

adjacent monomer units or amino acids, temperature factor of the theta atom, temperature factor of the iota atom, accessible surface area of the theta atom, accessible surface area of the iota atom, the number of different iota atom contacts for the particular theta atom and the number of other theta atoms on the same monomer unit as the theta atom.

In an embodiment, the 3-D coordinates of the contact pair and covalently attached adjacent atoms are normalized, as a group, to a common database reference frame as described below. This simplifies subsequent analysis of potential underlying contact patterns or rules and application of any such rules to drug design.

In an embodiment, the theta atom type is identified as being one and only one of: the 167 covalent atom types (excluding hydrogen) that make up the 20 natural amino acid building blocks of proteins (in this case the secondary database may be divided accordingly and comprise information about up to 27889, 167×167 , different contact types); and/or the 82 non-hydrogen atoms present in the 4 nucleotides of the deoxyribonucleic acid polymer (DNA); and/or the 42 non-hydrogen atoms present in the methylated DNA nucleotides, cytidine phosphate and adenosine phosphate; and/or the 85 non-hydrogen atoms present in the 4 nucleotide phosphates of the ribonucleic acid polymer (RNA); and/or the 89 non-hydrogen atoms present in 2-O'-methylated ribose nucleotide phosphates of RNA; and/or the over 400 non-hydrogen atoms present in the commonest post-transcription base modified RNA.

In an embodiment, the iota atom type is identified as being one and only one of: the 167 covalent atom types (excluding hydrogen) that make up the 20 natural amino acid building blocks of proteins; and/or the oxygen atom present in protein bound, structurally relevant, water molecules (this may be useful because crystal structures in the primary database often contain structurally relevant water molecules, i.e. certain protein atoms show definite interactions with bound water molecules); and/or the 82 non-hydrogen atoms present in the 4 nucleotides of the deoxyribonucleic acid polymer (DNA); and/or the 42 non-hydrogen atoms present in the methylated DNA nucleotides, cytidine phosphate and adenosine phosphate; and/or the 85 non-hydrogen atoms present in the 4 nucleotide phosphates of the ribonucleic acid polymer (RNA); and/or the 89 non-hydrogen atoms present in 2-O'-methylated ribose nucleotide phosphates of RNA; and/or the over 400 non-hydrogen atoms present in the commonest post-transcription base modified RNA.

In a contact pair the opposing atom is viewed and recorded from either side on the

contact. The nomenclature in an example embodiment is described below and illustrated schematically in Figure 1.

The atom on the reference side of the contact is termed the theta atom 1 whilst the opposing atom is termed the iota atom 2. In this example, the further atoms used for normalizing the 3-D coordinates are defined as follows. The next atom to which the theta atom 1 is covalently bonded, in the direction of the C alpha atom of that amino acid, is referred to as the third atom 3 and the next atom again, the fourth atom 4. The fourth atom 4 is covalently bonded to the third atom 3. The next atom to which the iota atom 2 is covalently bonded, in the direction of the C alpha atom of the respective amino acid, is termed the fifth atom 5 and the next again atom, the sixth atom 6. The sixth atom is covalently bonding to either the fifth atom or the iota atom 2.

In an embodiment, to avoid instances of ambiguity the third and fourth atoms are chosen uniquely for each specified theta atom type. In an embodiment, the fifth and sixth atoms are also chosen uniquely. In an embodiment, the following convention is applied. If the theta atom 1 happens to be a C alpha atom, then the third and fourth atoms are the backbone carbonyl carbon and oxygen atoms respectively. If the theta atom 1 is a backbone carbonyl carbon, then the third atom 3 and the fourth atom 4 are the C alpha carbon and the backbone nitrogen respectively. If the theta atom 1 is the backbone nitrogen, then the third atom 3 and the fourth atom 4 are the C alpha carbon and the backbone carbonyl carbon respectively. If the theta atom 1 is a C beta carbon atom, then the third atom 3 and the fourth atom 4 are the C alpha carbon and the backbone carbonyl carbon respectively. In phenylalanine and tyrosine side chains where there is a choice of two epsilon carbon atoms for the third and fourth atom positions, then the atom closest to the backbone nitrogen atom is selected.

In an embodiment, coordinate normalisation of each contact is performed on the theta, iota, third and fourth atoms, optionally also the fifth and sixth atoms, as a group so that their 3-D relationship is maintained. The resulting normalized coordinates may be referred to as a normalized coordinate group. In an embodiment, this is achieved by carrying out the following steps in sequence, as illustrated in Figures 2-6.

Figure 2 illustrates a theta atom 1, third atom 3 and fourth atom 4 of a non-bonding intra-molecular or inter-molecular contact positioned relative to a reference frame, defined relative to x-, y- and z-axes, according to coordinates given in a primary database (such as the wwPDB). In a first step of an example coordinate normalization process, the atom group coordinates are translated so that the theta atom 1 lies at the

zero coordinate (Figure 3). Next, the group as a whole is rotated about the z-axis until the third atom 3 is at $y = 0$ (Figure 4). Next, the group is rotated about the y-axis until the third atom is at $y=0$ and $z=0$ (Figure 5). Next, the group is rotated about the x-axis until the fourth atom 4 is at $y=0$ and the group as a whole lies in the x-y plane (all three atoms at $y=0$; Figure 6). In this manner each of the 167 first atom types can be superimposed for that type and the secondary database sub-divided accordingly. In turn each of the first atom divisions can be sub-divided into 167 iota atom types, facilitating the analysis of the spatial distribution of each iota atom type relative to each theta atom type.

In an embodiment, the distribution patterns of iota atoms relative to theta atoms are analysed in order to identify similarities between the distribution patterns for nominally different iota atom types. In this way, the unique iota atom types (e.g. the 167 covalent atom types mentioned above) can be combined into a number of groups (herein referred to as “predetermined groups of related iota atom types”) to simplify subsequent use of the data. Grouping together the atom types according to the similarity of distribution patterns reduces the computational load associated with the method described below with reference to Figure 7 for example, thus increasing speed and/or reducing hardware expense.

In an embodiment, this process is simplified by using polar coordinates rather than Cartesian coordinates (in an embodiment, this is achieved by performing conversion processing between Cartesian coordinates and polar coordinates, for example where the data in the primary database is presented using Cartesian coordinates). In an embodiment, two-dimensional polar coordinates are used, specifying the relative positions of the theta and iota atoms in terms only of the two polar angles θ (theta) and Φ (phi) (corresponding to latitude and longitude on a globe). The resulting two-dimensional latitude-longitude plots do not show any information about variations in the distance between the theta and iota atoms. However, it is found that this distance is relatively constant, so that the theta-phi plots contain most of the relevant information concerning the contact. Reducing the analysis to a problem in two dimensions rather than three greatly improves the efficiency of subsequent analyses. In an embodiment, contour lines are used to illustrate variations in the relative position of the iota atom. The contour lines may represent lines of constant “density” or probability of a relative positioning of the theta and iota atoms.

Analysis of such polar angle plots has revealed that a particularly important factor

governing the pattern of iota atom frequencies is the elemental nature and hybridisation state of the iota atoms, i.e. C sp³ , C sp²(aromatic), C sp²(non-aromatic), N sp³ , N sp², O sp³ , O sp² or S. As a result, it is possible to improve analysis efficiency by grouping the 167 atom types according to these identified eight groups. In other
5 embodiments, a different grouping may be used.

In general, environmental factors around the contact, such as the nature of adjacent amino acids, make less difference to the iota frequency pattern, with the exception of secondary structure. As might be expected the frequency patterns of backbone amide nitrogen theta atoms versus backbone oxygen iota atoms and vice-versa are skewed by
10 secondary structure, in particular as regards whether or not they are from beta sheet.

In an embodiment, the secondary database tags contact data with the local secondary structure type (helix, beta sheet or random coil). This provides the basis for differentiating any potential influence of secondary structure on contact patterns at a later stage.

15 In an embodiment, a method is provided based on the above that assists with the identification of modifications to a ligand that improve the strength of binding, or affinity, of the ligand to a binding site. In an embodiment, the method is used to assist with NCE or biologic drug design. In respect of the former, the method may be useful for predicting 'hotspots' or pharmacophore atom positions in potential drug binding
20 sites of target proteins. This can facilitate de novo drug design. In situations where there is an available structure of chemical matter bound in a binding site, the method can suggest atom types and positions for elaboration of the chemistry to obtain a ligand with better binding characteristics. In the case of protein drugs such antibodies, the method may be used to predict mutations in the protein or antibody binding site that
25 would lead to improvement in binding affinity or specificity. The method may also be used to suggest positions for modification within a macromolecular structure to improve the properties of the macromolecule. For example, as illustrated in the Examples section below, the method may be used to identify point mutations within antibody VH and VL chains in order to improve the thermal stability of the antibody. The mutations
30 are on separate chains, but are still within the antibody macromolecule.

Figure 7 illustrates an example method for designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target.

In step S1, data representing the target binding site of a target protein is obtained, for

example from a local or remote memory device 5. A target list of atoms forming the surface of the target binding site is identified.

In step S2, each atom in the target list is identified as a particular theta atom type.

In step S3, information is extracted from a structural database of biological
5 macromolecules (e.g. the wwPDB), provided for example by a local or remote memory device 7, about non-bonding, intra-molecular or inter-molecular contacts in which the first atom in a contacting pair of atoms is a particular theta atom type and the opposing, second atom of the pair is a particular iota atom type. The extracted information comprises spatial and/or contextual data about the iota atom relative to the theta atom.
10 The data is collected for a plurality of contacts of the given theta atom type and the resulting set of data is referred to as a theta contact set. In an embodiment, the theta contact set comprises data collected for all of the available contacts of the given theta atom type. The extracted information may form a database that is an example of the “secondary database” discussed above. In an embodiment, the information extracted in
15 step S3 is collected in a secondary database that comprises one and only one theta contact set for each of the theta atom types. In an example of such an embodiment, the theta contact sets of the secondary database are subdivided into a plurality of non-overlapping iota atom types or non-overlapping groups of related iota atom types. In an example of such an embodiment, the database is sub-divided to form groups of
20 identified contacts in which the first atom is one and only one of the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins and the second atom is in one and only one of a plurality of non-overlapping groups obtained by sorting the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins into groups based on chemical similarity.

25 In step S4, for each theta atom identified in the target list in step S2, data relating to a given iota atom type, or a predetermined group of related iota atom types, from the corresponding theta contact set extracted in step S3 is superimposed in or around the target binding site. In an embodiment, the superimposition comprises: parsing the theta contact set to extract spatial data for contacts comprising the given iota atom type or one
30 or more of the predetermined group of related iota atom types; and plotting this spatial data to determine theoretical locations representing where each iota atom type, or each of the one or more of the predetermined group of related iota atom types, would be located if: i) the theta atom of the contact were located at the position of the corresponding theta atom in the target binding site; and ii) the third and fourth atoms of

the contact were located at the positions of the third and fourth atoms of the corresponding theta atom in the target binding site. Where determined theoretical locations conflict with binding site atoms and/or are buried within the target protein, these may be removed from further analysis. For example if it is determined that the theoretical location of an individual iota atom intersects with the location of an atom of the target macromolecule closer than $R_w - 0.2$ angstroms then the iota atom is excluded from subsequent analysis.

In step S5, the superimposed data is combined and/or parsed in such a way as to predict one or more favoured regions of the binding site where the given iota type, or the predetermined group of related iota atom types, has high theoretical propensity.

In step S6, a candidate ligand is notionally docked into the binding site. Data defining the candidate ligand may be provided for example from a local or remote memory device 9. A comparison is then made between the type and position of one or more of the atoms of the candidate ligand with the predicted favoured regions for the respective iota atom types. On the basis of the comparison, modifications to the candidate ligand, in terms of alternate or additional candidate ligand atoms, are identified that will produce a greater intersection between the alternate and/or additional candidate ligand atoms and the respective iota atom type favoured regions, leading to an improvement in the affinity of the modified candidate ligand to the binding site compared to the unmodified candidate ligand.

In step S7, the modified candidate ligand is output either as a proposed improvement to an existing ligand or as part of an *ab initio* design of a new ligand. Optionally steps S7 and S6 can be iterated to further modify the ligand. The local or remote memory devices 5, 7 and 9 may be implemented in a single piece of hardware (e.g. a single storage device) or in two or more different, separate devices.

In an embodiment, the modified candidate ligand is output to an output memory device for storage or transmission and/or to a display for visualization.

In an embodiment, the type of a given theta atom is identified uniquely in step S2 such that there is no intersection between the group of contacts for which information is extracted in step S3 for the given theta atom and the group of contacts for any other theta atom type (with the exception of contacts involving the given theta atom type as the iota atom).

In an embodiment, step S5 comprises determining one or more favoured regions for each of a plurality of different iota atom types and/or predetermined groups of related

iota atom types. In such an embodiment, the comparison step S6 may be repeated for each of the plurality of different iota atom types and/or predetermined groups of related iota atom types, in order to identify potential modifications that involve the different iota atoms types or predetermined groups of related iota atom types.

5 In an embodiment, steps S2-S7 are performed for a plurality of different atoms in the binding site. In an embodiment, as described below, favoured regions may be determined more accurately by cumulatively combining (e.g. summing) the distributions of determined theoretical locations of the iota atom types as derived for a plurality of different atoms in the binding site.

10 In an embodiment, the analysis is extended such that, for each favoured region, vectors are derived that describe the position of the fifth atom relative to its respective iota atom. Analysis is carried out on the vectors to identify a favoured bond vector representing a prediction of the covalent attachment of a theoretical consensus iota atom in the region. The identified favoured bond vector can then be used to refine the design
15 of the candidate ligand and/or to refine the modification of the candidate ligand, as applicable. The identified favoured bond vector may be used for example to indicate how iota atoms in different favoured regions might be bonded together, thus assisting with the identification of modifications involving plural additions or exchanges of atoms. In an embodiment the analysis is a cluster analysis.

20 The distribution of theoretical locations gives a measure or propensity of how a particular iota atom type (or predetermined group of related iota atom types) will be favoured at different locations in the binding site. In an embodiment, a region in which a density of the theoretical locations is above a predetermined threshold is identified as one of the favoured regions. The density of theoretical locations is a measure of the
25 number of determined theoretical locations that occur in a given spatial volume for example. In an embodiment, iota atom theoretical locations are determined for a plurality of target atoms in the binding site and a region in which a density of the cumulative theoretical locations for the iota atom (or predetermined group of related iota atom types) for the plurality of target atoms is above the predetermined threshold is
30 identified as one of the favoured regions. The theoretical locations determined for different target atoms in the binding site may be summed, for example, in order to obtain the cumulative theoretical locations. This approach to taking into account the effects of different atoms in the binding site is computationally efficient and minimizes loss of information about the interaction between the candidate ligand and the binding

site. The approach is facilitated by the characterization of contacts in terms of pairs of simple atom types or simple atom types in combination with atoms of predetermined groups of simple atom types. Such an approach is not valid when contacts are characterized in terms of 3-atom fragments, such as is the case in Laskowski et al. for example.

The obtained distributions of theoretical locations can be transformed in various ways to create probability density functions, i.e. a statistical potential for the preference of a given atom type at a given position in the binding site. In turn, probability density functions can be treated in an analogous way to electron density and converted into ccp4 files which are a standard way of visualising such maps within molecular graphics software, e.g. Pymol.

In an embodiment, in step S5, the one or more favoured regions is/are expressed in polar coordinates, optionally comprising only the polar and azimuthal angles, optionally wherein the reference frame is normalized by reference to the third and fourth atoms 3,4.

In an embodiment, step S6 comprises: identifying a modification of the candidate ligand that increases a degree of overlap between an atom of the candidate ligand (whether present before the modification or not) and a predicted favoured region for an atom of the same type in the binding site. In an embodiment, the generated distributions of theoretical locations and/or favoured regions are inspected, for example by computer software or manually, as superimpositions on the target macromolecular structure in complex with the respective candidate ligand. If for instance the candidate ligand relates to an antibody, the interface between the antibody and target macromolecule may be examined to determine the degree of overlap between antibody atoms and the respective atom theoretical location distributions and/or favoured regions identified for that atom. In some cases the degree of overlap will already be high. However, in other regions of the interface the overlap may be low. It is in these regions where mutations in the adjacent amino acid residue of the antibody may be most effectively identified / proposed. In an embodiment each of the 19 other natural amino acids is considered in turn at this position, in all of their respective rotamer conformations and in each case the degree of overlap with the relevant atom theoretical location distributions and/or favoured regions is examined, with the aim of selecting those residues with the maximum degree of overlap for proposed mutations. In this manner, a rational means of selecting mutations is provided that may generate

affinity improvements in the chosen antibody. Individual point mutations in different regions of the antibody – target protein interface may be generated; those that lead to affinity improvement can be tested in combinations of two or more that may give synergistic increases in affinity.

5 In an embodiment, step S6 comprises replacing each of one or more of (optionally all of) the amino acid residues of the ligand that is/are in direct contact with the target binding site, or in close proximity to the target binding site, with each of one or more of (optionally all of) the residues chosen from the other 19 natural amino acids. Each such replacement, is referred to herein as a “residue replacement” and involves the
10 modification of a ligand by a single replacement of one residue with a different residue. In an embodiment, for each such residue replacement that does not cause conflict with adjacent atoms of the ligand or target (e.g overlap between one or more atoms of the replacement residue and one or more other atoms of the target or ligand), the type and position of each atom of the replacement residue is compared with the respective
15 atom type favoured regions to identify whether they will produce a greater intersection than the atoms of the original residue. In an embodiment a list is then output of the residue replacements that are identified as producing a greater intersection than atoms of the original residue. In an embodiment, for each of the listed residue replacements, the candidate ligand is then mutated to produce a modified, single-residue-mutated ligand
20 that incorporates the residue replacement. The affinity of each modified ligand to the target binding site can then be tested by experiment in order to identify those modifications which provide the greatest affinity improvement for the candidate ligand. For example, a group of residue replacements may be identified that yield a residue replacement that is greater than a predetermined threshold. In an embodiment, the
25 predetermined threshold may be zero so that the selected group consists only of residue replacements that improve the affinity to some extent. More advanced modifications of the candidate ligand can then be carried out based on this information. For example, in an embodiment the candidate ligand may be modified to incorporate a plurality of residue replacements, for example a plurality of those residue replacements that,
30 individually, were determined as providing the greatest affinity improvements. In this way it is possible to design a ligand that has an affinity that is improved even more than is possible by replacing only a single residue.

In some embodiments, lists of residue replacements may be produced that satisfy providing a greater intersection than atoms of the original residue (have a Δ IOTAScore

of less than zero). The lists may additionally be ranked based on other criteria. For example, lists may also be filtered based on $\Delta\Delta G$ scores (see below). Residue replacements with $\Delta\Delta G$ scores of less than zero imply stronger interactions compared with the original residue. Therefore a list may be produced where residues satisfy both
5 criteria of a $\Delta IOTAScore$ of less than zero (a negative $\Delta IOTAScore$), and a $\Delta\Delta G$ of less than zero (a negative $\Delta\Delta G$). This is illustrated in Example 2 below.

In the case of chemical matter, for instance a crystal structure of a low molecular weight chemical fragment bound in a pocket of a target protein, iota atom theoretical location distributions and/or favoured regions displayed in the binding site may suggest
10 atom types and vectors of chemical bonds for fragment growth that may yield a prototype NCE of higher potency.

In an embodiment, a plurality of modifications to the candidate ligand are identified. In this case, the method may further comprise selecting a subset of the identified modifications, for example to identify the modifications which are likely to be most
15 effective in terms of improving affinity. The selection may be carried out based on the extent to which the intersection between the alternate and/or additional candidate ligand atoms and the respective iota atom type favoured regions is greater compared to the unmodified candidate ligand. For example, modifications that result in an increase in the intersection that is above a predetermined threshold may be selected and
20 modifications that result in an increase in the intersection that is below a predetermined threshold may be discarded. An example of such a selection process is discussed below in the context of "Example 2". The " $\Delta IOTAScore$ " is an example of a measure of the extent to which the intersection between the alternate and/or additional candidate ligand atoms and the respective iota atom type favoured regions is greater compared to the
25 unmodified candidate ligand. Alternatively or additionally, the selection may be carried out based on the extent to which one or more factors contributing to the total energy of the complex formed by the binding of the modified candidate ligand to the binding site is/are reduced compared to the case where the unmodified candidate ligand is bound. For example, modifications that result in a decrease in the one or more factors (e.g. a
30 decrease in a sum of the one or more factors) that is above a predetermined threshold may be selected and modifications that result in a decrease in the one or more factors (e.g. a decrease in a sum of the one or more factors) that is below a predetermined threshold may be discarded. An example of such a selection process is discussed below in the context of "Example 2". The "Rosetta $\Delta\Delta G$ score" is an example of a measure of

the extent to which one or more factors contributing to the total energy of the complex formed by the binding of the candidate ligand to the binding site is/are reduced.

Examples of factors contributing to the total energy of the complex include a Lennard-Jones term, an implicit solvation term, an orientation-dependent hydrogen bond term, sidechain and backbone torsion potentials derived from the PDB, a short-ranged knowledge-based electrostatic term, and reference energies for each of the 20 amino acids that model the unfolded state, as discussed below.

In an embodiment, the method of identifying a modification to a candidate ligand is a computer-implemented method. In an embodiment, any one or more of the steps S1-S7 is/are performed on a computer. In an embodiment, all of the steps S1-S7 is/are performed on a computer. In addition, any one or more of the steps S101-S109 of Figure 14 (illustrating a workflow for predicting point mutations at the VH-VL interface of an antibody) may be automated. Any one or more of the steps S101-S109 may be performed on a computer. In one embodiment, all of the steps S101-S109 are automated. All of the steps S101-S109 may be performed on a computer.

A wide range of standard computing configurations, well known to the person skilled in the art, could be used as platforms to implement the method. The method is not limited to any particular hardware configuration, operating system or means for storing or transmitting software necessary for defining and/or implementing the method steps. In an embodiment, a computer readable medium or signal is provided that comprises computer readable instructions (e.g. code in a computer programming language) for causing a computer to carry out the method.

In an embodiment, a method of manufacturing a therapeutic ligand is provided. In an embodiment, the method of manufacturing comprises designing a new ligand or modifying an existing ligand according to one or more of the embodiments described above.

EXAMPLE 1: Affinity maturation of a Fab fragment of an anti-IL17F antibody

Introduction

In vitro methods of antibody affinity maturation are well known (see US 8 303 953 B2 column 13 lines 19 to 33). In a recent example Fujino *et al* (Fujino *et al* (2012) “Robust *in vitro* affinity maturation strategy based on interface-focused high-throughput mutational scanning”, Biochem. Biophys. Res. Comm., 428, 395-400)

report a high throughput mutational scanning strategy based on ribosome display panning of single point mutant single-chain Fab libraries at each of 50 identified antigen interface residues of the antibody, followed combinatorial ribosome display of enhanced binders that resulted in identification of a Fab with over 2000-fold affinity improvement. Such methods require a large investment in laboratory based resources and therefore various groups (reviewed by Kuroda *et al* (2012) “Computer-aided antibody design”, Prot. Eng. Design & Selection, 25, 507-521) have investigated *in silico* methods that predict improvements in antibody affinity so as to reduce or eliminate the need for screening large numbers of mutated antibody variants for improved affinity. These computer-aided antibody design protocols are either knowledge-based; i.e. using statistical potentials derived from observational data or physics-based, i.e. using and energy functions derived from models of the underlying physical interactions. Lippow *et al* (2007) (Lippow *et al* (2007), “Computational design of antibody-affinity improvement beyond *in vivo* maturation”, Nat Biotech., 25, 1171-1176) have achieved moderate success with the latter approach based on electrostatic interactions, but our understanding of parameterisation of such methods is still far from complete. Knowledge-based methods to date tend to identify individual antibody residues for random mutagenesis (e.g. Barderas *et al* (2008) “Affinity maturation of antibodies assisted by *in silico* modelling”, PNAS, 105, 9029-9034), which still entail considerable laboratory based effort.

In this Example, a knowledge-based approach was applied to affinity mature a Fab fragment of the anti-IL17F antibody described in US 8 303 953 B2. The final affinity matured antibody is described in WO 2012/095662 A1 as a full length IgG1 molecule. However, the method by which this antibody was affinity matured is not disclosed in the latter publication.

Methods

Identification of target (theta) atoms comprising the IL17F epitope

Using the coordinates of the co-crystal IL17F / Fab 496 complex structure described in WO 2009/130459 A2, all IL17F atoms within 6 Å of any Fab 496 atom were identified as epitope atoms and are listed in Table 1. Of this list of 209 theta atoms there are 86 specific theta atom types.

Table 1 :List of atoms comprising the IL17F epitope where the notation (1)-(2)-(3)-(4) designates (1) the respective F and I chains of the IL17 homodimer, (2) the amino acid residue and (3) residue number and (4) the atom type.

5

F-ASN-53-CG	F-ASN-89-C	F-ILE-129-CA	I-VAL-38-N
F-ASN-53-OD1	F-ASN-89-CA	F-ILE-129-CB	I-VAL-38-O
F-ALA-70-C	F-ASN-89-CB	F-ILE-129-CD1	I-SER-39-C
F-ALA-70-O	F-ASN-89-CG	F-ILE-129-CG1	I-SER-39-CA
F-GLN-71-C	F-ASN-89-N	F-ILE-129-CG2	I-SER-39-CB
F-GLN-71-CA	F-ASN-89-ND2	F-ILE-129-N	I-SER-39-N
F-GLN-71-CB	F-ASN-89-OD1	F-ILE-129-O	I-SER-39-OG
F-GLN-71-CD	F-SER-90-C	F-HIS-130-C	I-MET-40-C
F-GLN-71-CG	F-SER-90-CA	F-HIS-130-N	I-MET-40-N
F-GLN-71-N	F-SER-90-CB	F-HIS-130-O	I-MET-40-O
F-GLN-71-NE2	F-SER-90-N	F-HIS-131-C	I-SER-41-C
F-GLN-71-O	F-SER-90-O	F-HIS-131-CA	I-SER-41-CA
F-GLN-71-OE1	F-SER-90-OG	F-HIS-131-CE1	I-SER-41-CB
F-CYS-72-C	F-VAL-91-C	F-HIS-131-N	I-SER-41-N
F-CYS-72-CA	F-VAL-91-CA	F-HIS-131-ND1	I-SER-41-O
F-CYS-72-CB	F-VAL-91-CB	F-HIS-131-O	I-SER-41-OG
F-CYS-72-N	F-VAL-91-CG1	F-VAL-132-C	I-ARG-42-C
F-CYS-72-O	F-VAL-91-CG2	F-VAL-132-CA	I-ARG-42-CA
F-CYS-72-SG	F-VAL-91-N	F-VAL-132-CB	I-ARG-42-CB
F-ARG-73-C	F-VAL-91-O	F-VAL-132-CG1	I-ARG-42-CD
F-ARG-73-CA	F-PRO-92-C	F-VAL-132-CG2	I-ARG-42-CG
F-ARG-73-CB	F-PRO-92-CA	F-VAL-132-N	I-ARG-42-CZ
F-ARG-73-CG	F-PRO-92-CB	F-VAL-132-O	I-ARG-42-N
F-ARG-73-N	F-PRO-92-CD	F-GLN-133-C	I-ARG-42-NE
F-ARG-73-O	F-PRO-92-CG	F-GLN-133-CA	I-ARG-42-NH1
F-ASN-74-C	F-PRO-92-N	F-GLN-133-CB	I-ARG-42-NH2
F-ASN-74-CA	F-PRO-92-O	F-GLN-133-CD	I-ARG-42-O
F-ASN-74-CB	F-GLN-94-CD	F-GLN-133-CG	I-ASN-43-N
F-ASN-74-CG	F-GLN-94-NE2	F-GLN-133-N	I-ASN-43-OD1
F-ASN-74-N	F-GLN-94-OE1	F-GLN-133-O	I-ILE-44-CA
F-ASN-74-ND2	F-GLU-114-OE1	F-GLN-133-OE1	I-ILE-44-CB
F-ASN-74-O	F-LEU-117-CB	F-GLN-133-OXT	I-ILE-44-CD1
F-LEU-75-C	F-LEU-117-CD1	I-ILE-32-CG2	I-ILE-44-CG1
F-LEU-75-CA	F-LEU-117-CD2	I-ASN-33-CB	I-ILE-44-CG2
F-LEU-75-CB	F-LEU-117-CG	I-ASN-33-CG	I-ARG-47-CD
F-LEU-75-CD1	F-THR-119-CB	I-ASN-33-ND2	I-ARG-47-CZ
F-LEU-75-CD2	F-THR-119-CG2	I-ASN-33-OD1	I-ARG-47-NE
F-LEU-75-CG	F-THR-119-OG1	I-GLN-36-C	I-ARG-47-NH1
F-LEU-75-N	F-VAL-125-CB	I-GLN-36-CA	I-ARG-47-NH
F-LEU-75-O	F-VAL-125-CG1	I-GLN-36-CB	
F-GLU-84-OE1	F-VAL-125-CG2	I-GLN-36-CD	
F-ILE-86-C	F-PRO-127-C	I-GLN-36-NE2	
F-ILE-86-CA	F-PRO-127-CA	I-GLN-36-OE1	
F-ILE-86-CB	F-PRO-127-CB	I-ARG-37-C	
F-ILE-86-CD1	F-PRO-127-CD	I-ARG-37-CA	
F-ILE-86-CG1	F-PRO-127-CG	I-ARG-37-CB	
F-ILE-86-CG2	F-PRO-127-N	I-ARG-37-CD	
F-ILE-86-O	F-PRO-127-O	I-ARG-37-CG	
F-SER-87-C	F-VAL-128-C	I-ARG-37-CZ	
F-SER-87-CA	F-VAL-128-CA	I-ARG-37-N	
F-SER-87-N	F-VAL-128-CB	I-ARG-37-NE	
F-SER-87-O	F-VAL-128-CG1	I-ARG-37-NH1	
F-MET-88-C	F-VAL-128-CG2	I-ARG-37-NH2	
F-MET-88-CA	F-VAL-128-N	I-ARG-37-O	
F-MET-88-N	F-VAL-128-O	I-VAL-38-C	
F-MET-88-O	F-ILE-129-C	I-VAL-38-CA	

Creation of IOTA database

A secondary database of over 11 million intra-molecular atomic contact data was extracted from over 20000 non-homologous protein structures where resolution was ≤ 2 Å. Contacts were defined as any two atoms on opposing sides of a protein fold separated by a distance of $1\text{Å} +$ the sum of their respective Van de Waals radii or less, and were limited to atoms on residues at least 4 residues apart on the linear peptide sequence. The first atom of the contacting pair was designated the theta atom and the second atom, the iota atom. The database was divided into 167 contact sets according to the theta atom type, there being 167 non-hydrogen atom types within the 20 natural amino acid residues comprising proteins. Within a contact set the relative coordinates of each iota atom position was recorded after normalisation of the theta – iota atom pair coordinates. The latter was achieved by setting the theta atom to $x,y,z = 0,0,0$; the next covalently attached atom (3rd atom) to the theta atom (in the direction of the peptide backbone) to $x,y,z = x',0,0$ and the next again covalently attached atom (4th atom) to $x,y,z = x'',0,z'$. A consistent convention was employed to defined 3rd and 4th atoms.

Each theta contact set was further sub-divided into 167 iota atom types, but for convenience these were concatenated into 26 sub-groups according to chemical type based on the definition of Engh and Huber (Engh and Huber (1991) “Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement”, Acta Cryst., A47, 392-400).

In this Example the following iota sub-groups were employed:

Description	Atoms
Carbonyl O	all backbone O, Asn O δ 1, Gln O ϵ 1
Tetrahedral CH ₂	all C β 9except Ala, Ile, Thr, Val) Arg C δ ,C γ , Gln C γ , Glu C γ , Ile C γ 1, Lys C γ , C δ , C ϵ , Met C γ
Tetrahedral CH ₃	Ala C β , Ile C δ 1, C γ 2, Leu C δ 1, C δ 2, Met C ϵ , Thr C γ 2, Val C γ 1, C γ 2
NH	all backbone N (except Pro), Arg N ϵ , His N δ 1,N ϵ 2, Trp N ϵ 1
Hydroxyl O	Ser O γ , Thr O γ 1, Tyr O η
Carboxyl O	Asp O δ 1, O δ 2, Glu O ϵ 1, O ϵ 2
NH ₂	Asn N δ 2, Gln N ϵ 2

Superimposition of iota data over the IL17F epitope surface

For each of the 209 theta atoms comprising the IL17F epitope, the corresponding theta contact set was selected from the IOTA database and from that, an appropriate iota sub-

group was selected e.g. carbonyl oxygen. The relative iota coordinates from this sub-group were transposed relative to the reference frame of the given theta atom of the IL17F epitope (Table 2 illustrates example data). An iota dataset for a given sub-group was thus accumulated over the whole IL17F epitope. In cases where the location of a
5 given iota data point intersected with an atom of IL17F, closer than the sum of their respective Van de Waals radii minus $0.2\text{\AA}$, then these data points were excluded from the dataset. The process was repeated for all relevant iota sub-groups to produce a series of iota datasets for the IL17F epitope.

Table 2:

Id	out01:0000098	out01:0000127	out01:0000249	out01:0000280	out01:0000282	out01:0000283	out01:0000346	out01:0000366	out01:0000527	out01:0000542
pdcode	16vpA	16vpA	1a0cA	1a0pA	1a0pA	1a0pA	1a0sP	1a0sP	1a12A	1a12A
res	2.1	2.1	2.5	2.5	2.5	2.5	2.5	2.4	2.4	1.7
rval	0.26	0.26	0.177	0.287	0.287	0.287	0.228	0.228	0.219	0.219
org	Herpes simplex	Herpes simplex	Thermoanaerobacter	Escherichia coli	Escherichia coli	Escherichia coli	Salmonella typhimurium	Salmonella typhimurium	Homo sapiens	Homo sapiens
Tchain	A	A	A	A	A	A	P	P	A	A
Taalnd	96	235	36	126	137	137	85	138	257	12
Taanum	142	281	36	145	156	156	155	208	277	32
Taaname	THR	THR	THR	THR	THR	THR	THR	THR	THR	THR
Taass	C	E	E	H	E	E	H	E	C	C
Taaphi	-135.7	-129.6	-71.7	-87.1	-105.2	-105.2	-118.5	-123.9	-114.7	-127.1
Taapsi	163.3	160	157.7	-34.9	174	174	18.9	132.4	1	-10.7
Tomg	-179.5	178.5	177.5	178.7	177.6	177.6	-178.3	179.3	-178.5	-171.7
Tc1	61.3	52.2	66.6	60.5	64.4	64.4	53.1	-54.5	49.2	59.2
Tc2	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9
Tc3	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9
Tc4	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9
TUpAA	ALA	LEU	LYS	ALA	LEU	LEU	SER	GLY	GLY	SER
TDnAA	ARG	VAL	MET	GLY	MET	MET	SER	GLY	GLU	GLU
Tnum	724	1867	288	1032	1111	1111	626	1048	1919	93
Tname	C	C	C	C	C	C	C	C	C	C
Tbval	23.52	24.07	23.92	31.8	38.02	38.02	16.02	15.07	20.11	12.34
Tasa	0	0	0	0	0	0	1.84	0.19	0	0.21
Tcdist	0	0	0	0	0	0	0	0	0	0
Todist	1.234	1.233	1.24	1.215	1.239	1.239	1.226	1.222	1.247	1.249
Tndist	2.463	2.461	2.432	2.482	2.455	2.455	2.536	2.412	2.492	2.502
Tcadist	1.525	1.532	1.527	1.525	1.53	1.53	1.541	1.519	1.516	1.521
Icdist	3.472	3.069	3.359	3.618	3.647	3.199	3.565	3.589	3.432	3.67
Iodist	3.677	3.072	4.523	3.286	3.657	4.381	2.796	2.512	3.207	3.652
Indist	4.507	4.556	4.306	4.708	4.178	4.537	5.735	3.963	4.235	4.639
Itidist	3.472	3.069	3.359	3.618	3.647	3.199	3.565	3.589	3.432	3.67
Inum	1644	1696	244	2089	1535	1519	550	1213	1845	372
Iname	CD2	OD1	O	CD1	CB	O	NE2	N	O	CG1
Ichain	A	A	A	A	A	A	P	P	A	A
Iaalnd	206	213	30	258	190	188	75	159	247	50
Iaanum	252	259	30	279	209	207	145	229	267	70
Iaaname	LEU	ASP	GLU	TYR	LEU	ASP	GLN	GLY	SER	VAL
Iaass	H	E	C	H	C	C	E	E	C	E
Iaaphi	-94.9	-110.9	-60.4	-69.5	-73.2	-83.2	-125.6	-109.2	-149.7	-101.1
Iaapsi	-34.2	127.4	132.7	-18.6	-32.6	-14.4	142.7	141.1	38.2	-21
Iomg	-173.5	-178.2	-179.6	-177.8	-177.1	-179.8	177.8	178.2	-179.1	179.5
Ic1	-63.9	-171.8	176.2	-79.4	167.9	66	-66.4	999.9	-164	-65.2
Ic2	167.9	57.6	172.1	-57.1	69.8	-21.2	-63.5	999.9	999.9	999.9
Ic3	999.9	999.9	7	999.9	999.9	999.9	-52.7	999.9	999.9	999.9
Ic4	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9
IUpAA	ASP	CYS	GLU	ILE	VAL	ILE	GLY	TYR	LEU	VAL
IDnAA	PHE	LEU	VAL	THR	PHE	VAL	THR	ARG	ASN	GLN
Ibval	6.09	32.79	27.18	37.22	8.29	43.39	11.84	20.76	9.22	8.96
Iasa	0.13	0	0	0	0	0	0.69	0.25	0	1.04
K1cdist	4.988	4.114	4.557	4.099	4.209	4.342	4.034	4.42	3.962	4.187
K1odist	5.113	3.819	5.728	3.526	4.6	5.553	3.425	3.283	3.395	4.575
K1ndist	5.733	5.402	5.198	5.704	4.874	5.334	5.781	4.513	5.257	4.346
K1tdist	4.988	4.114	4.557	4.099	4.209	4.342	4.034	4.42	3.962	4.187
K1num	1642	1695	243	2088	1532	1518	548	1214	1844	371
K1name	CG	CG	C	CG	CA	C	CD	CA	C	CB
K1aanum	252	259	30	279	209	207	145	229	267	70
K1aaname	LEU	ASP	GLU	TYR	LEU	ASP	GLN	GLY	SER	VAL
K2cdist	NoValue	NoValue	NoValue	3.857	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2odist	NoValue	NoValue	NoValue	3.951	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2ndist	NoValue	NoValue	NoValue	4.353	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2tdist	NoValue	NoValue	NoValue	3.857	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2num	NoValue	NoValue	NoValue	2091	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2name	NoValue	NoValue	NoValue	CE1	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2aanum	NoValue	NoValue	NoValue	279	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2aaname	NoValue	NoValue	NoValue	TYR	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
Cx	0	0	0	0	0	0	0	0	0	0
Cy	0	0	0	0	0	0	0	0	0	0
Cz	0	0	0	0	0	0	0	0	0	0
CsphR	0	0	0	0	0	0	0	0	0	0
CsphT	0	0	0	0	0	0	0	0	0	0
CsphP	-3.142	0	0	3.142	0	0	0	0	0	0

Table 2 (continued):

id	out01:0000098	out01:0000127	out01:0000249	out01:0000280	out01:0000282	out01:0000283	out01:0000346	out01:0000366	out01:0000527	out01:0000542
CcylR	0	0	0	0	0	0	0	0	0	0
CcylT	-3.142	0	0	3.142	0	0	0	0	0	0
CcylZ	0	0	0	0	0	0	0	0	0	0
Ox	-0.628	-0.629	-0.648	-0.659	-0.666	-0.666	-0.609	-0.619	-0.609	-0.582
Oy	0.998	1.004	0.978	-0.754	1.028	1.028	-1.009	0.713	-1.084	-1.095
Oz	0.364	0.341	0.401	-0.689	0.182	0.182	0.341	0.777	0.085	-0.146
OsphR	1.234	1.233	1.24	1.216	1.238	1.238	1.227	1.223	1.246	1.249
OsphT	1.271	1.291	1.241	2.173	1.423	1.423	1.289	0.882	1.503	1.688
OsphP	2.132	2.13	2.156	-2.289	2.146	2.146	-2.114	2.286	-2.083	-2.059
OcylR	1.234	1.233	1.24	1.216	1.238	1.238	1.227	1.223	1.246	1.249
OcylT	2.132	2.13	2.156	-2.289	2.146	2.146	-2.114	2.286	-2.083	-2.059
OcylZ	0.364	0.341	0.401	-0.689	0.182	0.182	0.341	0.777	0.085	-0.146
Nx	2.055	2.054	2.006	2.096	2.039	2.039	2.158	1.982	2.11	2.128
Ny	1.357	1.356	1.375	1.328	1.368	1.368	1.332	1.374	1.327	1.316
Nz	0	0	0	0	0	0	0	0	0	0
NsphR	2.463	2.461	2.432	2.481	2.455	2.455	2.536	2.412	2.493	2.502
NsphT	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571
NsphP	0.584	0.583	0.601	0.565	0.591	0.591	0.553	0.606	0.561	0.554
NcylR	2.463	2.461	2.432	2.481	2.455	2.455	2.536	2.412	2.493	2.502
NcylT	0.584	0.583	0.601	0.565	0.591	0.591	0.553	0.606	0.561	0.554
NcylZ	0	0	0	0	0	0	0	0	0	0
Cax	1.525	1.532	1.527	1.525	1.53	1.53	1.541	1.519	1.516	1.521
Cay	0	0	0	0	0	0	0	0	0	0
Caz	0	0	0	0	0	0	0	0	0	0
CasphR	1.525	1.532	1.527	1.525	1.53	1.53	1.541	1.519	1.516	1.521
CasphT	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571
CasphP	0	0	0	0	0	0	0	0	0	0
CacylR	1.525	1.532	1.527	1.525	1.53	1.53	1.541	1.519	1.516	1.521
CacylT	0	0	0	0	0	0	0	0	0	0
CacylZ	0	0	0	0	0	0	0	0	0	0
Tx	-12.209	3.16	75.678	17.53	11.229	11.229	-34.605	-58.479	39.484	5.281
Ty	-17.125	-7.179	45.43	-33.077	-23.148	-23.148	-6.201	-0.92	-28.232	-11.189
Tz	-3.618	-14.31	44.154	38.22	26.483	26.483	-24.486	-14.58	0.224	9.116
TsphR										
TsphT										
TsphP										
TcylR										
TcylT										
TcylZ										
Ix	-0.948	-0.804	1.174	-1.669	-0.297	0.852	-2.689	-1.363	1.533	-0.362
Iy	0.629	-0.729	-2.203	1.536	1.127	-2.85	-0.806	3.054	-2.417	-0.092
Iz	-3.281	2.871	-2.247	-2.818	-3.455	-1.176	2.198	1.302	-1.893	-3.651
IsphR	3.473	3.069	3.359	3.617	3.646	3.199	3.565	3.589	3.432	3.67
IsphT	2.808	0.361	2.304	2.464	2.816	1.947	0.907	1.2	2.155	3.04
IsphP	2.556	-2.405	-1.081	2.398	1.828	-1.28	-2.85	1.991	-1.006	-2.893
IcylR	3.473	3.069	3.359	3.617	3.646	3.199	3.565	3.589	3.432	3.67
IcylT	2.556	-2.405	-1.081	2.398	1.828	-1.28	-2.85	1.991	-1.006	-2.893
IcylZ	-3.281	2.871	-2.247	-2.818	-3.455	-1.176	2.198	1.302	-1.893	-3.651
K1x	-1.194	-1.35	1.899	-2.979	0.085	1.65	-2.134	-1.06	0.644	0.446
K1y	1.1	-0.243	-2.896	1.098	-0.13	-3.76	-0.567	3.344	-3.182	1.142
K1z	-4.717	3.878	-2.962	-2.593	-4.206	-1.413	3.376	2.69	-2.271	-4.003
K1sphR	4.989	4.113	4.557	4.099	4.209	4.342	4.034	4.421	3.962	4.187
K1sphT	2.81	0.34	2.278	2.256	3.105	1.902	0.579	0.917	2.181	2.844
K1sphP	2.397	-2.963	-0.99	2.788	-0.992	-1.157	-2.882	1.878	-1.371	1.198
K1cylR	4.989	4.113	4.557	4.099	4.209	4.342	4.034	4.421	3.962	4.187
K1cylT	2.397	-2.963	-0.99	2.788	-0.992	-1.157	-2.882	1.878	-1.371	1.198
K1cylZ	-4.717	3.878	-2.962	-2.593	-4.206	-1.413	3.376	2.69	-2.271	-4.003
K2x	NoValue	NoValue	NoValue	-1.258	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2y	NoValue	NoValue	NoValue	2.771	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2z	NoValue	NoValue	NoValue	-2.37	NoValue	NoValue	NoValue	NoValue	NoValue	NoValue
K2sphR				3.857						
K2sphT				2.232						
K2sphP				1.997						
K2cylR				3.857						
K2cylT				1.997						
K2cylZ				-2.37						
inter	MS	MS	MM	MS	MS	MM	MS	MM	MM	MS
Icount	1	6	3	4	4	4	2	4	4	2
Tcount	1	1	1	1	2	2	1	1	1	1

Table 2 : Key	
Header	Description
Id	Incremented Identifier of a contact. Every unique identifier is a unique theta-iota interaction. The id is eight digits with place-holding 0s (ex. 1 is 00000001).
pdrcode	The PDB code that was given in the input file. (ex 1mu4B for chain B of 1mu4, or 1mu4 for the structure).
res	The resolution of the PDB structure.
rval	The R-value of the PDB structure.
org	The organism source for the PDB structure.
Tchain	The chain identifier for the theta atom.
Taaind	The amino acid index for the theta amino acid (index starts at 1 for the first amino acid in the structure and increments for each amino acid)
Taanum	The amino acid number for the theta atom.
Taaname	The amino acid name for the theta atom.
Taass	The amino acid secondary structure for the theta atom.
Taaphi	The amino acid phi angle for the theta atom.
Taapsi	The amino acid psi angle for the theta atom.
Tomg	The amino acid omega angle for the theta atom.
Tchi1	The amino acid chi 1 angle for the theta atom.
Tchi2	The amino acid chi 2 angle for the theta atom.
Tchi3	The amino acid chi 3 angle for the theta atom.
Tchi4	The amino acid chi 4 angle for the theta atom.
TUpAA	The amino acid which is upstream of the theta amino acid.
TDnAA	The amino acid which is downstream of the theta amino acid.
Tnum	The atom number for the theta atom.
Tname	The atom name for the theta atom.
Tbval	The B value or temperature factor for the theta atom.
Tasa	The Accessible Surface Area of the theta atom.
Tcdist	The distance from the theta atom to the backbone Carbon.
Todist	The distance from the theta atom to the backbone Oxygen.
Tndist	The distance from the theta atom to the backbone Nitrogen.
Tcadist	The distance from the theta atom to the backbone Alpha Carbon.
Icdist	The distance from the iota atom to the backbone Carbon of the theta amino acid.
Iodist	The distance from the iota atom to the backbone Oxygen of the theta amino acid.
Indist	The distance from the iota atom to the backbone Nitrogen of the theta amino acid.
Itdist	The distance from the iota atom to the theta atom.
Inum	The atom number for the iota atom.
Iname	The atom name for the iota atom.
Ichain	The chain identifier for the iota atom.
Iaaind	The amino acid index for the iota amino acid (index starts at 1 for the first amino acid in the structure and increments for each amino acid)

Table 2 : Key (continued)	
Header	Description
Ia anum	The amino acid number for the iota atom.
Ia name	The amino acid name for the iota atom.
Ia ass	The amino acid secondary structure for the iota atom.
Ia phi	The amino acid phi angle for the iota atom.
Ia psi	The amino acid psi angle for the iota atom.
I omg	The amino acid omega angle for the iota atom.
I chi1	The amino acid chi 1 angle for the iota atom.
I chi2	The amino acid chi 2 angle for the iota atom.
I chi3	The amino acid chi 3 angle for the iota atom.
I chi4	The amino acid chi 4 angle for the iota atom.
I UpAA	The amino acid which is upstream of the iota amino acid.
I DnAA	The amino acid which is downstream of the iota amino acid.
I bval	The B value or temperature factor for the iota atom.
I asa	The Accessible Surface Area for the iota atom.
K1 cdist	The distance from the first kappa atom to the backbone Carbon of the theta amino acid.
K1 odist	The distance from the first kappa atom to the backbone Oxygen of the theta amino acid.
K1 ndist	The distance from the first kappa atom to the backbone Nitrogen of the theta amino acid.
K1 tdist	The distance from the first kappa atom to the theta atom.
K1 num	The atom number for the first kappa atom.
K1 name	The atom name for the first kappa atom.
K1 a anum	The amino acid number for the first kappa atom.
K1 a name	The amino acid name for the first kappa atom.
K2 cdist	The distance from the second kappa atom to the backbone Carbon of the theta amino acid.
K2 odist	The distance from the second kappa atom to the backbone Oxygen of the theta amino acid.
K2 ndist	The distance from the second kappa atom to the backbone Nitrogen of the theta amino acid.
K2 tdist	The distance from the second kappa atom to the theta atom.
K2 num	The atom number for the second kappa atom.
K2 name	The atom name for the second kappa atom.
K2 a anum	The amino acid number for the second kappa atom.
K2 a name	The amino acid name for the second kappa atom.
Cx	The Theta-superimposed x coordinate for the backbone Carbon atom of the theta AA.
Cy	The Theta-superimposed y coordinate for the backbone Carbon atom of the theta AA.
Cz	The Theta-superimposed z coordinate for the backbone Carbon atom of the theta AA.

Table 2 : Key (continued)	
Header	Description
CsphR	The Theta-superimposed spherical polar distance for the backbone Carbon of the theta AA.
CsphT	The Theta-superimposed spherical polar latitude angle for the backbone Carbon of the theta AA.
CsphP	The Theta-superimposed spherical polar longitude angle for the backbone Carbon of the theta AA.
CcylR	The Theta-superimposed cylindrical polar distance for the backbone Carbon of the theta AA.
CcylT	The Theta-superimposed cylindrical polar angle for the backbone Carbon of the theta AA.
CcylZ	The Theta-superimposed cylindrical polar z coordinate for the backbone Carbon of the theta AA.
Ox	The Theta-superimposed x coordinate for the backbone Oxygen atom of the theta AA.
Oy	The Theta-superimposed y coordinate for the backbone Oxygen atom of the theta AA.
Oz	The Theta-superimposed z coordinate for the backbone Oxygen atom of the theta AA.
OsphR	The Theta-superimposed spherical polar distance for the backbone Oxygen of the theta AA.
OsphT	The Theta-superimposed spherical polar latitude angle for the backbone Oxygen of the theta AA.
OsphP	The Theta-superimposed spherical polar longitude angle for the backbone Oxygen of the theta AA.
OcylR	The Theta-superimposed cylindrical polar distance for the backbone Oxygen of the theta AA.
OcylT	The Theta-superimposed cylindrical polar angle for the backbone Oxygen of the theta AA.
OcylZ	The Theta-superimposed cylindrical polar z coordinate for the backbone Oxygen of the theta AA.
Nx	The Theta-superimposed x coordinate for the backbone Nitrogen atom of the theta AA.
Ny	The Theta-superimposed y coordinate for the backbone Nitrogen atom of the theta AA.
Nz	The Theta-superimposed z coordinate for the backbone Nitrogen atom of the theta AA.
NsphR	The Theta-superimposed spherical polar distance for the backbone Nitrogen of the theta AA.
NsphT	The Theta-superimposed spherical polar latitude angle for the backbone Nitrogen of the theta AA.
NsphP	The Theta-superimposed spherical polar longitude angle for the backbone Nitrogen of the theta AA.
NcylR	The Theta-superimposed cylindrical polar distance for the backbone Nitrogen of the theta AA.

Table 2 : Key (continued)	
Header	Description
NcylT	The Theta-superimposed cylindrical polar angle for the backbone Nitrogen of the theta AA.
NcylZ	The Theta-superimposed cylindrical polar z coordinate for the backbone Nitrogen of the theta AA.
Cax	The Theta-superimposed x coordinate for the Alpha Carbon atom of the theta AA.
Cay	The Theta-superimposed y coordinate for the Alpha Carbon atom of the theta AA.
Caz	The Theta-superimposed z coordinate for the Alpha Carbon atom of the theta AA.
CasphR	The Theta-superimposed spherical polar distance for the Alpha Carbon of the theta AA.
CasphT	The Theta-superimposed spherical polar latitude angle for the Alpha Carbon of the theta AA.
CasphP	The Theta-superimposed spherical polar longitude angle for the Alpha Carbon of the theta AA.
CacylR	The Theta-superimposed cylindrical polar distance for the Alpha Carbon of the theta AA.
CacylT	The Theta-superimposed cylindrical polar angle for the Alpha Carbon of the theta AA.
CacylZ	The Theta-superimposed cylindrical polar z coordinate for the Alpha Carbon of the theta AA.
Tx	The Theta-superimposed x coordinate for the theta atom.
Ty	The Theta-superimposed y coordinate for the theta atom.
Tz	The Theta-superimposed z coordinate for the theta atom.
TsphR	The Theta-superimposed spherical polar distance for the Theta atom.
TsphT	The Theta-superimposed spherical polar latitude angle for the Theta atom.
TsphP	The Theta-superimposed spherical polar longitude angle for the Theta atom.
TcylR	The Theta-superimposed cylindrical polar distance for the Theta atom.
TcylT	The Theta-superimposed cylindrical polar angle for the Theta atom.
TcylZ	The Theta-superimposed cylindrical polar z coordinate for the Theta atom.
Ix	The Theta-superimposed x coordinate for the iota atom.
Iy	The Theta-superimposed y coordinate for the iota atom.
Iz	The Theta-superimposed z coordinate for the iota atom.
IsphR	The Theta-superimposed spherical polar distance for the iota atom.
IsphT	The Theta-superimposed spherical polar latitude angle for the iota atom.
IsphP	The Theta-superimposed spherical polar longitude angle for the iota atom.
IcylR	The Theta-superimposed cylindrical polar distance for the iota atom.

Table 2 : Key (continued)	
Header	Description
IcylT	The Theta-superimposed cylindrical polar angle for the iota atom.
IcylZ	The Theta-superimposed cylindrical polar z coordinate for the iota atom.
K1x	The Theta-superimposed x coordinate for the first kappa atom.
K1y	The Theta-superimposed y coordinate for the first kappa atom.
K1z	The Theta-superimposed z coordinate for the first kappa atom.
K1sphR	The Theta-superimposed spherical polar distance for the first kappa atom.
K1sphT	The Theta-superimposed spherical polar latitude angle for the first kappa atom.
K1sphP	The Theta-superimposed spherical polar longitude angle for the first kappa atom.
K1cylR	The Theta-superimposed cylindrical polar distance for the first kappa atom.
K1cylT	The Theta-superimposed cylindrical polar angle for the first kappa atom.
K1cylZ	The Theta-superimposed cylindrical polar z coordinate for the first kappa atom.
K2x	The Theta-superimposed x coordinate for the second kappa atom.
K2y	The Theta-superimposed y coordinate for the second kappa atom.
K2z	The Theta-superimposed z coordinate for the second kappa atom.
K2sphR	The Theta-superimposed spherical polar distance for the second kappa atom.
K2sphT	The Theta-superimposed spherical polar latitude angle for the second kappa atom.
K2sphP	The Theta-superimposed spherical polar longitude angle for the second kappa atom.
K2cylR	The Theta-superimposed cylindrical polar distance for the second kappa atom.
K2cylT	The Theta-superimposed cylindrical polar angle for the second kappa atom.
K2cylZ	The Theta-superimposed cylindrical polar z coordinate for the second kappa atom.
inter	The type of interaction for the theta & iota contact: MM – Main chain to Main chain, MS – Main chain to Side chain, SM..., SS... where the first letter refers to theta and the second refers to iota.
Icount	The iota atom count for the particular theta atom.
Tcount	The theta atom count for the particular amino acid.

Visualisation of IL17F iota datasets

Each iota dataset was visualized in relation to the IL17F / Fab 496 structure using molecular graphics computer software such as Pymol. This could be done by direct plotting of the iota dataset as individual points or by first mathematically transforming the dataset into a density function and a file format compatible for molecular graphic

display e.g. ccp4, so that contour maps of higher density could be displayed over the IL17 epitope.

Inspection of iota density maps for intersection with Fab 496 paratope atoms

5 The IL17F / Fab 496 interface was examined to determine the degree of intersection between individual Fab 496 atoms per residue and the corresponding iota density maps. Residues were identified where there was no or little intersection. In these cases alternative residues were substituted via the molecular graphics software to determine whether better intersection could be achieved between residue atoms and relevant iota
10 density maps. Amino acid substitutions producing good iota density map intersection were short listed for *in vitro* production and testing as single point mutations as intact IgG versions of Fab 496.

DNA manipulations and general methods

15 *E. coli* strain INVαF (Invitrogen) was used for transformation and routine culture growth. DNA restriction and modification enzymes were obtained from Roche Diagnostics Ltd. and New England Biolabs. Plasmid preparations were performed using Maxi Plasmid purification kits (Qiagen, catalogue No. 12165). DNA sequencing reactions were performed using ABI Prism Big Dye terminator sequencing kit
20 (catalogue No. 4304149) and run on an ABI 3100 automated sequencer (Applied Biosystems). Data was analysed using the program Auto Assembler (Applied Biosystems). Oligonucleotides were obtained from Invitrogen. The concentration of IgG was determined by IgG assembly ELISA.

25 *Affinity maturation of antibody CA028_00496*

CA028_0496 is a humanised neutralising antibody which binds both IL17A and IL17F isoforms. It comprises the grafted variable regions, termed gL7 and gH9, whose sequences are disclosed in WO 2008/047134. The wild type Fab' fragment of this antibody (Fab 496) and mutant variants were prepared as follows: oligonucleotide
30 primer sequences were designed and constructed in order to introduce single point mutations in the light chain variable region (gL7) as per residues and positions determined in the above short list. Each mutated light chain was separately sub-cloned into the UCB Celltech human light chain expression vector pKH10.1, which contained DNA encoding the human C-kappa constant region (Km3 allotype). The unaltered
35 heavy chain variable region (gH9) sequence was sub-cloned into the UCB Celltech

expression vector pVhg1Fab6His which contained DNA encoding human heavy chain gamma-1 constant region, CH1. Heavy and light chain encoding plasmids were co-transfected into HEK293 cells using the 293fectin™ procedure according to the manufacturer's instructions (Invitrogen, Catalogue No. 12347-019). IgG1 antibody levels secreted into the culture supernatants after 10 to 12 days culture were assessed by ELISA and binding kinetics assessed by surface plasmon resonance (see below). Mutants showing improved or similar binding to IL17F were then prepared and tested in combination as double, triple, quadruple or quintuple light chain mutations as above.

Surface Plasmon Resonance (SPR)

All SPR experiments were carried out on a Biacore 3000 system (Biacore AB) at 25°C using HBS-EP running buffer (10 mM HEPES pH 7.4, 150 mM NaCl 3mM EDTA 0.005% (v/v) surfactant P20, Biacore AB). Goat F(ab')₂ anti-IgG Fab' specific antibody (Jackson Labs. Product code 109-006-097) was covalently attached to the surface of a CM5 sensor chip (GE Healthcare) by the amine coupling method, as recommended by the manufacturers. Briefly, the carboxymethyl dextran surface was activated with a fresh mixture of 50 mM N-hydroxysuccinimide and 200 mM 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide for 5 minutes at a flow rate of 10 µl / min. Anti-Fab antibody at 50 µg / ml in 10 mM sodium acetate pH 5.0 buffer was injected for 60 sec at the same flow rate. Finally the surface was deactivated with a 10 minute pulse of 1 M ethanolamine.HCl pH8.5, leaving of 4000 to 5000 response units (RU) of immobilized antibody on the chip. A reference flow cell was prepared on the same chip by omitting the protein from the above procedure.

Wild type and mutated 496 Fabs were harvested from culture supernatants in the range 3 to 30 µg / ml and crude supernatants were diluted in running buffer into the range 0.5 to 2 µg / ml. In order to evaluate binding kinetics to IL17F, each antibody was first captured on the anti-Fab' surface by injection at 10 µl / min for 60 sec to yield an additional 150 to 250 RU signal. Recombinant human IL17F was titrated from 10 nM in running buffer and injected at 30 µl / ml, to produce an association phase over 180 sec followed by a dissociation phase of 300 sec. At the end of each cycle the surface was regenerated with a 60 sec pulse of 40 mM HCl followed by a 30 sec pulse of 5 mM NaOH at 10 µl / min. For each Fab' a control cycle was carried out where the IL17F injection was replaced with an injection of running buffer.

Sensograms were corrected by subtraction of reference flow cell signal, then by subtracting the control cycle sensogram for the respective Fab'. Dissociation rate

constants (k_d) and association rate constants (k_a) were fitted to the data using Biaevaluation software (Biacore AB). Fab affinities (K_D) were calculated as $K_D = k_d / k_a$.

5 Results

Intersection of iota density maps with Fab 496 atoms

Inspection of that part of the surface of Fab 496 forming the interface with IL17F revealed that in many areas there was complete intersection between a given Fab 496
 10 atom and the corresponding iota density map. This was particularly the case for amino acid residues comprising the Fab 496 heavy chain. However there were a number of regions comprising the light chain where there was no intersection or little intersection of Fab 496 atoms and corresponding iota density maps (Panel A, Figs 8 to 13).
 By substituting alternative residues into the Fab 496 structure at these positions it was
 15 possible to achieve intersection between one or more atoms of the substituted residue and the corresponding iota density maps. (Panel B, Figs 8 to 13). Thus on Fab 496 light chain the threonine 30 C γ 2 atom does not intersect with the methyl carbon iota density map and the O γ 1 atom does not intersect with the hydroxyl oxygen iota density map. However when threonine 30 is mutated to arginine, there is intersection between the N ϵ
 20 atom and the secondary amide iota density map (Fig. 8).
 Figure 9 shows tetrahedral methylene iota density and secondary amide iota density dispersed around light chain arginine 54 without any intersection with respective side chain atoms of this residue. Conversely when mutated to a serine at this position there is good intersection of hydroxyl oxygen iota density and the gamma oxygen atom of the
 25 serine side chain. The light chain serine 56 gamma oxygen atom does not intersect with hydroxyl oxygen iota density. However when this residue is mutated to isoleucine there is intersection of the delta 1 methyl atom of the isoleucine side chain and methyl iota density (Fig. 10). In the case of serine 60, again the gamma oxygen is too distant from adjacent hydroxyl oxygen iota density to achieve intersection, (Fig. 11), but mutating
 30 this residue to aspartate brings about intersection of the side chain delta oxygen atom with carboxyl oxygen iota density.
 Light chain threonine 72 is not a CDR residue but part of framework 3, its side chain methyl atom does not intersect with methyl carbon iota density nor its side chain oxygen atom with hydroxyl oxygen iota density. But the arginine 72 mutation allows

intersection of both side chain delta carbon atom with methylene carbon iota density and of side chain eta nitrogen atoms with guanidinium nitrogen iota density.

Effects of iota designed mutations on antibody CA028_00496 on binding kinetics

- 5 Five iota designed single point light chain mutations in Fab 496, T30R, R54S, S56I, S60D and T72R, showed small improvements in binding affinity ranging from 1.7 to 3.6 fold. The improvement observed with the non-CDR mutation, T72R, was surprising since residues at this position do not normally contact the antigen. For all but one of these mutations (S60D), the improvement was driven by a reduction in dissociation rate
- 10 constant (Table 3). Combinations of these mutations in pairs resulted in a synergistic improvement in binding, with a 3.8 to 7.5 fold reduction in dissociation rate constant; with triple and quadruple combinations producing further step reductions in dissociation rate constant (Table 3).
- 15 The combination of all five light chain mutations produced the largest improvement in binding affinity (Table 4) to give an affinity value of 11 pM to IL17F, some 180-fold better than the original Fab 496. An important finding was that there was no deleterious effect on the binding to the IL17A isoform, in fact an improvement, with affinity constant at 2 pM compared to 14 pM for the original Fab 496. It is interesting that
- 20 combinations of the designed mutations produce a synergistic enhancement of affinity; one explanation is that they are reasonably spaced across the light chain paratope surface (Fig. 13) and therefore avoid negative interaction effects.

25 *Table 3 : Effect of single residue and combined residue light chain mutations of antibody CA028_000496 on IL17F dissociation rate constant*

CA028_00496 light chain mutations	k_d (s ⁻¹)	fold change
wt	4.2E-03	
T30R	3.2E-03	1.3
R54S	2.1E-03	2.0
S56I	3.0E-03	1.4
S60D	4.1E-03	1.0
T72R	2.5E-03	1.7
T30R / R54S	7.0E-04	5.8
T30R / S56I	1.1E-03	3.8
R54S / S56I	6.2E-04	6.6

R54S / T72R	5.5E-04	7.5
S56I / T72R	8.0E-04	5.1
S60D / T72R	1.2E-03	3.6
T30R / R54S / T72R	2.5E-04	19
T30R / S56I / T72R	4.1E-04	10
T30R / S56I / S60D / T72R	3.1E-04	14
T30R / R54S / S56I / T72R	1.3E-04	35
T30R / R54S / S60D / T72R	9.3E-05	44
T30R / R54S / S56I / S60D / T72R	5.4E-05	104

Table 4: Affinity constant of antibody CA028_000496 versus variant comprising 5 light chain mutations

	k_a ($M^{-1}s^{-1}$)	k_d (s^{-1})	K_D (M)	K_D (pM)
wild type	2.0E+06	4.1E-03	2.0E-09	2000
T30R / R54S / S56I / S60D / T72R	2.3E+06	2.6E-05	1.1E-11	11

Conclusion

The method of creating iota density maps over the epitope surface of IL17F in order to predict favourable mutations in the corresponding antibody, Fab 496, has proven to be successful in that the affinity constant has been improved 180-fold to a K_D of 11 pM. This method does not assume that only CDR mutations can be used but demonstrates that framework region mutations are also important for affinity maturation.

EXAMPLE 2: Utilisation of an automated process of the invention in the region between an Ab's heavy/light chains to improve stability

Methods

Identification of target (theta) atoms of both the heavy and light chains of the Fv interface of Fab X

Using the coordinates of the crystal structure of an antibody Fab fragment "Fab X" complexed with antigen, all heavy chains atoms within 6 Å of any light chain atoms were identified as epitope atoms (167 theta atoms listed in Table 5). Similarly, all light chains atoms within 6 Å of any heavy chain atoms were identified as epitope atoms (159 theta atoms listed in Table 5).

Table 5 :List of atoms comprising the heavy and light chain epitopes where the notation (1)-(2)-(3) designates (1) the respective H and L chains of Fab X, (2) residue number and (3) the atom type.

5 **Binding pocket (THETA) = H Binding partner (IOTA) = L**

H-47-CA	H-57-O	H-104-OH	H-117-N
H-47-CB	H-60-CB	H-109-CB	H-117-O
H-47-CG1	H-60-CD1	H-109-CG1	H-117-OH
H-47-CG2	H-60-CG1	H-109-CG2	H-118-C
H-49-CB	H-60-CG2	H-112-CD1	H-118-CA
H-49-CD	H-62-CE2	H-112-CE1	H-118-CB
H-49-CG	H-62-CH2	H-112-CE2	H-118-CD1
H-49-NE2	H-62-CZ2	H-112-CZ	H-118-CD2
H-49-OE1	H-62-NE1	H-112-OH	H-118-CE1
H-53-C	H-68-C	H-113-C	H-118-CE2
H-53-CA	H-68-CA	H-113-CA	H-118-CG
H-53-O	H-68-CB	H-113-CB	H-118-CZ
H-54-C	H-68-CD1	H-113-O	H-118-N
H-54-CA	H-68-CD2	H-113-OG	H-118-O
H-54-N	H-68-CE1	H-114-C	H-119-C
H-54-O	H-68-CE2	H-114-CA	H-119-CA
H-55-C	H-68-CG	H-114-CB	H-119-CB
H-55-CA	H-68-CZ	H-114-CG2	H-119-CG
H-55-CB	H-68-N	H-114-N	H-119-N
H-55-CD1	H-68-O	H-114-O	H-119-O
H-55-CD2	H-69-C	H-114-OG1	H-121-C
H-55-CG	H-69-CA	H-115-C	H-121-CA
H-55-N	H-69-N	H-115-CA	H-121-CB
H-55-O	H-69-O	H-115-CB	H-121-CD1
H-56-C	H-70-C	H-115-N	H-121-CD2
H-56-CA	H-70-CA	H-115-O	H-121-CE2
H-56-CB	H-70-CB	H-116-C	H-121-CE3
H-56-N	H-70-N	H-116-CA	H-121-CG
H-56-O	H-70-O	H-116-CB	H-121-CH2
H-57-C	H-71-C	H-116-CD	H-121-CZ2
H-57-CA	H-71-CA	H-116-CG	H-121-CZ3
H-57-CB	H-71-CB	H-116-N	H-121-N
H-57-CD1	H-71-CG2	H-116-O	H-121-NE1
H-57-CD2	H-71-N	H-117-C	H-121-O
H-57-CE2	H-71-OG1	H-117-CA	H-122-C
H-57-CE3	H-72-N	H-117-CB	H-122-CA
H-57-CG	H-104-CD1	H-117-CD1	H-122-N
H-57-CH2	H-104-CD2	H-117-CD2	H-122-O
H-57-CZ2	H-104-CE1	H-117-CE1	H-123-C
H-57-CZ3	H-104-CE2	H-117-CE2	H-123-CA
H-57-N	H-104-CG	H-117-CG	H-123-O
H-57-NE1	H-104-CZ	H-117-CZ	

Binding pocket (THETA) = L Binding partner (IOTA) = H

L-11-OD1	L-52-C	L-60-CZ	L-106-CG2
L-41-O	L-52-O	L-60-N	L-106-N
L-42-C	L-53-C	L-60-O	L-106-O
L-42-CA	L-53-CA	L-61-C	L-107-C
L-42-CB	L-53-N	L-61-CA	L-107-CA
L-42-O	L-53-O	L-61-CB	L-107-CB
L-43-C	L-54-C	L-61-CD	L-107-N
L-43-CA	L-54-CA	L-61-CG	L-107-O
L-43-CB	L-54-CB	L-61-N	L-107-OG
L-43-CD1	L-54-N	L-61-O	L-108-C
L-43-CD2	L-54-O	L-61-OE1	L-108-CA
L-43-CE1	L-55-C	L-61-OE2	L-108-CB
L-43-CE2	L-55-CA	L-62-N	L-108-CG
L-43-CG	L-55-CB	L-64-NZ	L-108-N
L-43-CZ	L-55-CD	L-66-CG2	L-108-OD1
L-43-N	L-55-CG	L-98-CB	L-108-OD2
L-43-O	L-55-N	L-98-CD1	L-109-C
L-44-C	L-55-O	L-98-CD2	L-109-CA
L-44-CA	L-56-C	L-98-CE1	L-109-CB
L-44-N	L-56-CA	L-98-CE2	L-109-CG2
L-44-O	L-56-CB	L-98-CG	L-109-N
L-45-C	L-56-N	L-98-CZ	L-109-O
L-45-CA	L-56-O	L-98-OH	L-109-OG1
L-45-CB	L-57-C	L-100-C	L-110-CA
L-45-N	L-57-CA	L-100-CA	L-110-N
L-45-O	L-57-CB	L-100-N	L-111-CA
L-45-OG	L-57-CD1	L-100-O	L-111-CB
L-47-CB	L-57-CD2	L-101-C	L-111-CD1
L-47-CD1	L-57-CG	L-101-CA	L-111-CD2
L-47-CD2	L-57-N	L-101-N	L-111-CE1
L-47-CE1	L-59-C	L-101-O	L-111-CE2
L-47-CE2	L-59-O	L-102-C	L-111-CG
L-47-CG	L-60-C	L-102-CA	L-111-CZ
L-47-CZ	L-60-CA	L-102-N	L-111-N
L-47-OH	L-60-CB	L-105-O	L-111-O
L-49-CB	L-60-CD1	L-106-C	L-112-O
L-49-CD	L-60-CD2	L-106-CA	L-113-C
L-49-CG	L-60-CE1	L-106-CB	L-113-CA
L-49-NE2	L-60-CE2	L-106-CD1	L-113-O
L-49-OE1	L-60-CG	L-106-CG1	

Superimposition of iota data over the heavy and light chain interface surface

For each of the 167 theta atoms comprising the heavy chain epitope, the corresponding

5 theta contact set was selected from the IOTA database and from that, an appropriate iota

sub-group was selected e.g. carbonyl oxygen. The relative iota coordinates from this sub-group were transposed relative to the reference frame of the given theta atom of the heavy chain epitope. An iota dataset for a given sub-group was thus accumulated over the whole heavy chain epitope. In cases where the location of a given iota data point intersected with an atom of the heavy chain, closer than the sum of their respective Van de Waals radii minus 0.2Å, then these data points were excluded from the dataset. The process was repeated for all relevant iota sub-groups to produce a series of iota datasets for the heavy chain epitope.

For each of the 159 theta atoms comprising the light chain epitope, the corresponding theta contact set was selected from the IOTA database and from that, an appropriate iota sub-group was selected e.g. carbonyl oxygen. The relative iota coordinates from this sub-group were transposed relative to the reference frame of the given theta atom of the heavy chain epitope. An iota dataset for a given sub-group was thus accumulated over the whole light chain epitope. In cases where the location of a given iota data point intersected with an atom of the light chain, closer than the sum of their respective Van de Waals radii minus 0.2Å, then these data points were excluded from the dataset. The process was repeated for all relevant iota sub-groups to produce a series of iota datasets for the light chain epitope.

Inspection of iota density maps for intersection with heavy-light chain atoms

The whole process was automatically performed with an internal customised Rosetta python library script tailored for mutable positions identification, single point mutant generation, low-energy rotamer state enumeration, quantitative IOTA score computation, VH-VL chains binding energy estimation, and point mutants prioritisation.

Two scoring methods were used for mutants ranking:

1. ΔIOTAScore

Aforementioned IOTA density maps generated were used to compute the spatial intersection values between each heavy atom of residue at each mutable position and the density critical points in the corresponding type of maps nearby. IOTAScore is the sum of the volumetric overlaps between the heavy atoms of one residue with the maximum of IOTA densities with the corresponding type definitions, which reflects the degree of intersection between individual Fab X atoms per residue and the corresponding iota density maps. IOTAScore is negative numerically, where lower values imply more

intersection. Δ IOTAScore is the change of IOTAScores between the mutant residue and the wildtype one; similarly, the more negative the Δ IOTAScore value the greater the implication that the mutant is more favoured than the wildtype one.

2. Rosetta $\Delta\Delta$ G score

- 5 The Rosetta energy function is a linear combination of terms that model interaction forces between atoms, solvation effects, and torsion energies. More specifically, Score12, the default full atom energy function in Rosetta is composed of a Lennard-Jones term, an implicit solvation term, an orientation-dependent hydrogen bond term, sidechain and backbone torsion potentials derived from the PDB, a short-ranged
- 10 knowledge-based electrostatic term, and reference energies for each of the 20 amino acids that model the unfolded state. The binding strength between two binding partners, or Δ G, can be computed by subtracting the Rosetta scores of the individual partners alone with that of the complex structure formed by the two partners. Lower Δ G implies stronger binding. $\Delta\Delta$ G is the change of Δ G between the mutant complex and the
- 15 wildtype one; the more negative the $\Delta\Delta$ G value the greater the implication that the mutant binding affinity is higher than the wildtype one.

Figure 14 illustrates the workflow for *in silico* predicting point mutation at the VH-VL interface of the Fab X structure.

- 20 In step S101, all residues on the heavy chain with at least one heavy atom within 8 Å of any light chain heavy atoms were identified as mutable positions. Similarly, all the residues on light chain with at least one heavy atom within 8 Å of any heavy chain heavy atoms were identified as mutable positions.
- In step S102, for the wildtype Fab X crystal structure, the residue-wise IOTAScores and
- 25 binding energy Δ G are computed, respectively. In step S102.1, the IOTAScore for the wildtype residue on the current mutable position with the corresponding IOTA density maps nearby is computed, termed as ($\text{IOTAScore}_{\text{wt}}, \text{Position}_j$); in step S102.2, the binding energy of wildtype Fab X VH and VL chains is computed with Rosetta score12 function, termed as ΔG_{wt} .
- 30 In step S103, the wildtype residue on the current mutable position identified in step S101 are replaced (mutated) by the other amino acid types. Out of the 20 natural amino acid types, proline and cysteine are excluded from mutation. All the other 18 types (alanine, arginine, asparagine, aspartic acid, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, serine, threonine, tryptophan,

tyrosine, valine) except the wildtype itself are mutated on each mutable position one by one.

In step S104, for each mutated residue type at each mutable position, the top 100 lowest-energy (in terms of Rosetta scoring function) rotamer states are generated using

5 Rosetta. The other high energy rotamer states are discarded.

In step S105, for each rotamer state of mutant residue generated in S104, the IOTAScore is computed in the same way of step S102, termed as (IOTAScore_{mutant}, Position_j, Type_k, Rotamer_i).

10 In step S106, the Δ IOTAScore for the current combination of rotamer state, mutant residue type, and mutable position is computed by subtracting (IOTAScore_{wt}, Position_j) with (IOTAScore_{mutant}, Position_j, Type_k, Rotamer_i), which is termed as (Δ IOTAScore, Position_j, Type_k, Rotamer_i). Steps S105 and S106 were repeated to compute all of the Δ IOTAScores for each rotamer states for the current mutant residue type and mutable position.

15 In step S107, the optimal rotamer state of the current mutant residue type and mutable position is determined with the lowest Δ IOTAScore value, as shown in step S107.1. The binding energy of the mutant with the optimal rotamer state is computed in step S107.2 in the same way as step S102.2, termed as (Δ G_{mutant}, Position_j, Type_k). In step S107.3, the change of binding energies $\Delta\Delta$ G between mutant and wildtype is calculated by subtraction of Δ G_{wt} with Δ G_{mutant}. After the optimal rotamer state is prioritised, steps 20 S103 to S107 were repeated for the next mutant amino acid type at the current mutable position.

In step S108, for the current mutable position, only the candidate mutants satisfying the criteria of both Δ IOTAScore < 0 and $\Delta\Delta$ G < 0 are kept for later ranking. The rest are 25 discarded. Steps S102 to S108 were repeated to go through all the mutable positions and generate all candidate mutants satisfying the same criteria.

In step S109, all the candidate mutant structures were outputted for later visualisation analysis. The final list of candidate mutants were sorted and ranked by the lowest Δ IOTAScores.

30

The running command and parameters used were as below:

For light chain mutations prediction, the command was:

“python multiRotamersFabInterfaceIOTAScan.py --pdb FabX.pdb --only_chains L --region all --useIOTA --IOTAtype 167 --output_mutant”

For heavy chain mutations prediction, the command was:

“python multiRotamersFabInterfaceIOTAScan.py --pdb FabX.pdb --only_chains H --region all --useIOTA --IOTAtype 167 --output_mutant”

Extra Rosetta relevant parameters were initialized by adding the following code to the

5 “multiRotamersFabInterfaceIOTAScan.py”:

“init(extra_options = “-ex1 -ex2 -score:weights score12 -no_his_his_pairE -constant_seed -edensity:mapreso 3.0 -correct -mute all”

DNA manipulations and general methods

10 *E. coli* strain INVαF (Invitrogen) was used for transformation and routine culture growth. DNA restriction and modification enzymes were obtained from Roche Diagnostics Ltd. and New England Biolabs. Plasmid preparations were performed using Maxi Plasmid purification kits (Qiagen, catalogue No. 12165). DNA sequencing reactions were performed using ABI Prism Big Dye terminator sequencing kit
15 (catalogue No. 4304149) and run on an ABI 3100 automated sequencer (Applied Biosystems). Data was analysed using the program Auto Assembler (Applied Biosystems). Oligonucleotides were obtained from Invitrogen. The concentration of IgG was determined by IgG assembly ELISA.

20 *Thermostability improvement of Fab X through affinity maturation of the heavy-light chain interface*

The wild type Fab fragment of Fab X and mutant variants were prepared as follows: oligonucleotide primer sequences were designed and constructed in order to introduce single point mutations in both the heavy and light chain variable regions as per residues
25 and positions determined in the above short list. Each mutated light chain was separately sub-cloned into the UCB Celltech human light chain expression vector pKH10.1, which contained DNA encoding the human C-kappa constant region (Km3 allotype). Each mutated heavy chain variable region sequence was separately sub-cloned into the UCB Celltech expression vector pVhg1Fab6His which contained DNA
30 encoding human heavy chain gamma-1 constant region, CH1. Heavy and light chain encoding plasmids were co-transfected into HEK293 cells using the 293fectin™ procedure according to the manufacturer’s instructions (Invitrogen. Catalogue No. 12347-019). IgG1 Fab antibody levels secreted into the culture supernatants after 10 to

12 days culture were assessed by ELISA and binding kinetics assessed by surface plasmon resonance (see below).

Mutants showing improved thermostability were then prepared and tested in combination as double, or triple mutations as above.

5

Surface Plasmon Resonance (SPR)

All SPR experiments were carried out on a BIAcore T200 (GE Healthcare). Affinipure F(ab')₂ Fragment goat anti-human IgG, F(ab')₂ fragment specific (Jackson ImmunoResearch) was immobilised on a CM5 Sensor Chip via amine coupling chemistry to a capture level of ≈ 5000 response units (RUs). HBS-EP buffer (10mM HEPES pH 7.4, 0.15 M NaCl, 3 mM EDTA, 0.05 % Surfactant P20, GE Healthcare) was used as the running buffer with a flow rate of 10 μ L/min. A 10 μ L injection of Fab X at 0.75 μ g/mL was used for capture by the immobilised anti-human IgG-F(ab')₂. Antigen was titrated over the captured Fab X at various concentrations (50nM to 6.25nM) at a flow rate of 30 μ L/min. The surface was regenerated by 2 x 10 μ L injection of 50 mM HCl, followed by a 5 μ L injection of 5 mM NaOH at a flowrate of 10 μ L/min. Background subtraction binding curves were analysed using the T200evaluation software (version 1.0) following standard procedures. Kinetic parameters were determined from the fitting algorithm.

20

Thermostability assay

Thermofluor assay was performed to assess the thermal stabilities of purified molecules. Purified proteins (0.1 mg/ml) were mixed with SYPRO® Orange dye (Invitrogen), and the mixture dispensed in quadruplicate into a 384 PCR optical well plate. Samples were analysed on a 7900HT Fast Real-Time PCR System (Agilent Technologies) over a temperature range from 20°C to 99°C, with a ramp rate of 1.1°C/min. Fluorescence intensity changes per well were plotted against temperature and the inflection points of the resulting slopes were used to generate the T_m.

25

Results

30

Intersection of iota density maps with heavy and light chain atoms of the interface

The automated method using a Rosetta scan produced a table of mutations ranked by

IOTA score (Table 6).

Effects of iota designed mutations on Fab X on thermostability

Six iota-designed single point mutations in Fab X, H-T71R, H-T71K, H-T71N, H-T71H, L-S107E and L-T109I, showed small improvements in thermostability ranging from 0.5°C to 2.9°C over wild-type (Table 7). Figures 15, 16 and 17 provide computer generated visualisations depicting the effects of these single point mutations. In particular, these Figures show that whilst H-T71, L-S107 and L-T109 have no density intersection, H-R71 intersects with the amide density, L-E107 intersects with the carboxylate density and L-I109 intersects with the methyl density. Combinations of these mutations in pairs resulted in a synergistic improvement in thermostability; with triple combinations producing further step improvements in thermostability (Table 8).

The combination of H-T71R, L-S107E and L-T109I, mutations produced the largest improvement in thermostability (Table 8) to give a T_m of 81.2°C, some 5.8°C better than the original Fab X. This combination of the three mutations is depicted in Figure 18. An important finding was that there was no significant loss on the binding to its antigen.

Table 6 : Proposed mutations generated by the Rosetta scan method and their ranking in order of IOTA score

Rank (by IOTAScore)	Heavy chain	$\Delta\Delta G$	$\Delta IOTA$ Score	Light chain	$\Delta\Delta G$	$\Delta IOTA$ Score
1	T71R	-0.37	-56.27	T109H	-1.49	-92.37
2	V109R	-0.04	-50.58	T109K	-0.01	-67.33
3	V109K	-0.51	-40.67	T109I	-1.69	-45.17
4	T71H	-1.06	-38.94	I106L	-2.11	-42.11
5	T71K	-0.08	-36.09	T109L	-1.58	-37.44
6	T71W	-0.67	-25.95	I106H	-3.5	-31.77
7	T71Y	-0.67	-25.64	S107E	-0.62	-18.71
8	T71N	-0.03	-24.89	A53Y	-0.93	-17.68

9	D119W	-0.89	-23.58	A53F	-1.13	-15.64
10	T71Q	-0.38	-23.28	I106N	-0.47	-13.57
11	V109I	-0.56	-22.64			
12	V109H	-0.4	-21.73			

Table 7 : Thermostability of mutations compared with wild-type thermostability of 75.7 °C

Rank	Heavy chain	$\Delta\Delta G$	$\Delta IOTA$ Score	Tm °C	Light chain	$\Delta\Delta G$	$\Delta IOTA$ Score	Tm °C
1	T71R	-0.37	-56.27	78.6	T109H	-1.49	-92.37	75.2
2	V109R	-0.04	-50.58	75.6	T109K	-0.01	-67.33	ND
3	V109K	-0.51	-40.67	75.6	T109I	-1.69	-45.17	79.2
4	T71H	-1.06	-38.94	76.6	I106L	-2.11	-42.11	74.5
5	T71K	-0.08	-36.09	78.6	T109L	-1.58	-37.44	77.6
6	T71W	-0.67	-25.95	75.6	I106H	-3.5	-31.77	72.2
7	T71Y	-0.67	-25.64	75.3	S107E	-0.62	-18.71	77.3
8	T71N	-0.03	-24.89	76.2	A53Y	-0.93	-17.68	68.8
9	D119W	-0.89	-23.58	73.7	A53F	-1.13	-15.64	67.2
10	T71Q	-0.38	-23.28	75.8	I106N	-0.47	-13.57	71.8
11	V109I	-0.56	-22.64	ND				
12	V109H	-0.4	-21.73	75.1				

5 ND=Not Determined

Table 8 : Thermostability and affinity of combinations of mutations.

Combination	Rosetta ddG (VH/VL)	Rosetta dE (Fv)	Tm (°C)	Tm SD	KD (nM)
-------------	---------------------	-----------------	---------	-------	---------

H-T71R + L-S107E + L-T109I	-3.72	-1.73	81.2	0.3	ND
H-T71R + L-T109I	-2.3	-1.17	80.1	0	0.88
H-T71R + L-S107E	-2.13	-0.93	80.1	0	3.90
H-T71K + L-S107E	-0.3	0.76	79.4	0.6	3.50
H-T71K + L-S107E + L-T109L	-4.11	-1.41	79.3	0.3	ND
H-T71K + L-T109I	-1.96	-0.98	79.1	0.3	1.39
H-T71R + L-T109L	-2.77	-0.95	78.0	0.6	0.96
H-T71K + L-T109L	-2.2	-0.7	77.7	0.1	1.59
H-T71R	-0.37		78.0	0.3	0.80
H-T71K	-0.08		77.9	0.3	1.27
L-S107E	-0.62		77.3	0.5	5.52
L-T109I	-1.69		77.9	0.3	1.47
L-T109L	-1.58		76.9	0.5	1.73
WT			75.4	0.2	1.35

ND=Not Determined

CLAIMS

1. A method for designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target, comprising:
 - a) identifying a target list of atoms forming the surface of the target binding site;
 - b) identifying each atom, hereinafter referred to as a theta atom, in the target list, as a particular theta atom type;
 - c) extracting from a structural database of biological macromolecules, information about non-bonding, intra-molecular or inter-molecular atom to atom contacts, where the first atom in a contacting pair of atoms is of a particular theta atom type and the opposing, second atom of the pair, hereinafter referred to as an iota atom, is of a particular iota atom type, said information comprising spatial and/or contextual data about the iota atom relative to the theta atom, and said data collected for a plurality of contacts of the given theta atom type from the said database is hereinafter referred to as a theta contact set;
 - d) for each theta atom identified in the target list in step b), superimposing in or around the target binding site data relating to a given iota atom type, or a predetermined group of related iota atom types, from the corresponding theta contact set extracted in step c);
 - e) combining and / or parsing the superimposed data in such a way as to predict one or more favoured regions of the binding site where the given iota atom type, or the predetermined group of related iota atom types, has high theoretical propensity; and
 - f) with a candidate ligand notionally docked into the binding site, comparing the type and position of one or more of the atoms of the candidate ligand with the predicted favoured regions for the respective iota atom types, to identify a modification to the candidate ligand, in terms of alternate and/or additional candidate ligand atoms, that will produce a greater intersection between the alternate and/or additional candidate ligand atoms and the respective iota atom type favoured regions, leading to an improvement in the affinity of the modified candidate ligand to the binding site compared to the unmodified candidate ligand;

wherein each non-bonding intra-molecular or inter-molecular contact in the database is defined as a contact between opposing residues of a protein fold or between opposing monomer units of a macromolecular fold or between two interacting macromolecular

partners and is specifically between a theta atom on one side of the fold or first interacting partner and an iota atom on the opposing side or second interacting partner; in an instance where the following condition is satisfied:

$s - R_w \leq t$, where s is the separation between the two atoms of the contact, R_w is the sum of the van de Waals radii of the two atoms of the contact, and t is a predetermined threshold distance; and

wherein the theta atom type is identified uniquely in step b) such that there is no intersection between the data of a theta contact set extracted in step c) for a given theta atom type and the data of any other theta contact set extracted in step c) for any other theta atom type, apart from data concerning contacts involving the given theta atom as the iota atom.

2. A method according to claim 1, wherein for each non-bonding intra-molecular contact extracted from the structural database of protein members in step c) the following condition is also satisfied:

the theta atom and the iota atom of the contact are on different residues separated by at least four residues along the linear polypeptide or are on separate polypeptide chains.

3. A method according to any of the preceding claims, wherein the theta atom type is identified as being one and only one of:

the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins;

the 82 non-hydrogen atoms present in the 4 nucleotides of the deoxyribonucleic acid polymer (DNA);

the 42 non-hydrogen atoms present in the methylated DNA nucleotides, cytidine phosphate and adenosine phosphate;

the 85 non-hydrogen atoms present in the 4 nucleotide phosphates of the ribonucleic acid polymer (RNA);

the 89 non-hydrogen atoms present in 2-O'-methylated ribose nucleotide phosphates of RNA;

the over 400 non-hydrogen atoms present in the commonest post-transcription base modified RNA.

4. A method according to any of the preceding claims, wherein the information extracted in step c) is collected in a secondary database comprising one and only one

theta contact set for each of the theta atom types.

5. A method according to claim 4, wherein each of said secondary database theta contact sets is sub-divided into a plurality of non-overlapping iota atom types or non-overlapping groups of related iota atom types.

6. A method according to any of the preceding claims, wherein the iota atom type is identified as being one and only one of:

- the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins;
- the oxygen atom present in protein bound, structurally relevant, water molecules;
- the 82 non-hydrogen atoms present in the 4 nucleotides of the deoxyribonucleic acid polymer (DNA);
- the 42 non-hydrogen atoms present in the methylated DNA nucleotides, cytidine phosphate and adenosine phosphate;
- the 85 non-hydrogen atoms present in the 4 nucleotide phosphates of the ribonucleic acid polymer (RNA);
- the 89 non-hydrogen atoms present in 2-O'-methylated ribose nucleotide phosphates of RNA; and/or
- the over 400 non-hydrogen atoms present in the commonest post-transcription modified bases of RNA.

7. A method according to any of the preceding claims, wherein said predetermined group of related iota atom types is one of a plurality of non-overlapping groups obtained by sorting the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins into groups of similar chemical type.

8. A method according to claim 7, wherein the iota atom types are sorted into the plurality of non-overlapping groups according to one or more of the following factors: elemental nature of the atom type, hybridisation state of the atom type.

9. A method according to claim 7 or 8, wherein the iota atom types are sorted into a plurality of non-overlapping groups comprising the following: C sp³, C sp²(aromatic), C sp²(non-aromatic), N sp³, N sp², O sp³, O sp², S.

10. A method according to any of the preceding claims, wherein:
- said spatial data extracted in step c) defines the position of each iota atom specified in the theta contact set by geometrical reference to the position of the theta atom and to the positions of third and fourth atoms;
 - the third atom is covalently bonded to the theta atom; and
 - the fourth atom is covalently bonded to the third atom.
11. A method according to claim 10, wherein:
- for each iota atom specified in the theta contact set, said spatial data extracted in step c) defines the position of fifth and sixth atoms by geometrical reference to the position of the theta atom and to the positions of the third and fourth atoms;
 - the fifth atom is covalently bonded to the iota atom; and
 - the sixth atom is covalently bonded to either the fifth atom or the iota atom.
12. A method according to claim 11, wherein the superimposition in or around the target site of step (d) comprises:
- parsing the theta contact set to extract spatial data for contacts comprising the given iota atom type or one or more of the predetermined group of related iota atom types; and
 - plotting this spatial data to determine theoretical locations representing where each iota atom type, or each of the one or more of the predetermined group of related iota atom types, would be located if: i) the theta atom of the contact were located at the position of the corresponding theta atom in the target binding site; and ii) the third and fourth atoms of the contact were located at the positions of the third and fourth atoms of the corresponding theta atom in the target binding site.
13. A method according to claim 12, wherein:
- the extracted spatial data is parsed against said contextual data before said plotting step.
14. A method according to claims 12 and 13, wherein a region in which a density of theoretical locations for the given iota atom type, or for the one or more of the predetermined group of related iota atom types, is above a predetermined threshold is identified as one of the favoured regions.

15. A method according to any of claims 12-14, wherein theoretical locations for the given iota atom type, or for one or more of the predetermined group of related iota atom types, are determined for a plurality of theta atoms on the target list and a region in which a density of the cumulative theoretical locations is above the predetermined threshold is identified as one of the favoured regions.
16. A method according to any of claims 12-15, wherein:
if the theoretical location of an individual iota atom intersects with the location of an atom of the target macromolecule closer than $R_w - 0.2$ angstroms then the said iota atom is excluded from subsequent analysis.
17. A method according to any of claims 11-16, wherein:
the third and fourth atoms are chosen uniquely for each specified theta atom type.
18. A method according to any of claims 11-17, wherein:
the fifth and sixth atoms are chosen uniquely for each specified iota atom type.
19. A method according to any of claims 11-18, wherein:
for each favoured region, vectors are derived to describe the position of the fifth atom relative to its respective iota atom and analysis is carried out on said vectors in order to identify a favoured bond vector representing a prediction of the covalent attachment of a theoretical consensus iota atom in the said region, said identified favoured bond vector being used to refine the design of the candidate ligand or modification of the candidate ligand.
20. A method according to any of the preceding claims, wherein:
said contextual data extracted in step (c), contains contextual information concerning the local environment of each contact pair in the theta contact set, including one or more of the following in any combination: secondary structure, amino acid types or other monomer types comprising the contact pair, adjacent monomer units and/or local geometry thereof in a polymer chain either side of the contact, adjacent amino acids in a polypeptide chain on either side of the contact, local geometry of the said adjacent monomer units or amino acids, temperature factor of the theta atom, temperature factor of the iota atom, accessible surface area of the theta atom, accessible surface area of the

iota atom, the number of different iota atom contacts for the particular theta atom and the number of other theta atoms on the same monomer unit as the theta atom.

21. A method according to any of the preceding claims, wherein step (f) comprises:
identifying a modification of the candidate ligand that increases a degree of overlap between one or more atoms of the candidate ligand and a predicted favoured region or regions for an iota atom type or predetermined group of related iota atom types in the binding site.
22. A method according to any of the preceding claims, wherein a plurality of modifications to the candidate ligand are identified in step (f) and the method further comprises selecting a subset of the identified modifications based on one or both of the following: 1) the extent to which the intersection between the alternate and/or additional candidate ligand atoms and the respective iota atom type favoured regions is greater compared to the unmodified candidate ligand; and 2) the extent to which one or more factors contributing to the total energy of the complex formed by the binding of the modified candidate ligand to the binding site is/are reduced compared to the case where the unmodified candidate ligand is bound.
23. A method according to any of the preceding claims, wherein $t = 2.5$ angstroms
24. A method according to any of the preceding claims, wherein $t = 0.8$ angstroms
25. A method according to any of the preceding claims, further comprising:
out-putting data representing the modification identified in step (f).
26. A method according to any of claims 1-25, wherein the ligand is a protein.
27. A method according to claim 26, wherein the ligand is an antibody.
28. A method according to claim 26 or 27, wherein step (f) comprises replacing each of one or more of the amino acid residues of the ligand that is/are in direct contact with the target binding site, or in close proximity to the target binding site, with each of one or more alternative residues chosen from the other 19 natural amino acids, each

replacement being referred to as a residue replacement,

wherein for each residue replacement that does not cause conflict between the replacement residue and adjacent atoms of the ligand or target, the type and position of each atom of the replacement residue is compared with the respective atom type favoured regions to identify whether they will produce a greater intersection than the atoms of the original residue.

29. A method according to claim 28, further comprising:

outputting a list of the residue replacements that are identified as producing a greater intersection than atoms of the original residue;

for each listed residue replacement, using mutation of the candidate ligand to produce a modified ligand that incorporates the residue replacement;

testing the affinity of each of the modified ligands to the target binding site in order to determine which residue replacements result in an affinity improvement that is above a predetermined threshold.

30. A method according to claim 29, further comprising:

modifying the candidate ligand to incorporate a plurality of the residue replacements that have been determined to result in an affinity improvement that is above the predetermined threshold.

31. A computer readable medium or signal comprising computer readable instructions for causing a computer to carry out the method of any one of the preceding claims.

32. A medium or signal according to claim 31, wherein the computer is caused to carry out at least steps (c) – (e).

33. A medium or signal according to claim 31 or 32, wherein the computer is caused to carry to carry out at least step (f).

34. A method of manufacturing a therapeutic ligand, comprising:

designing a therapeutic ligand according to the method of any of claims 1-30; and manufacturing the therapeutic ligand thus designed.

35. A therapeutic ligand manufactured according to the method of claim 34.
36. A method or ligand according to claim 34 or 35, wherein the ligand is a protein.
37. A method or ligand according to claim 36, wherein the protein is an antibody.
38. A method of generating a database for use in a method for designing a ligand *ab initio* that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target, comprising:
- analysing the relative positions of atoms in each of a plurality of proteins or other biological macromolecules in order to identify instances of a non-bonding intra-molecular contact between a first atom, referred to as a theta atom, and a second atom, referred to as an iota atom, of the protein or macromolecule; and
 - generating a database that for each identified contact specifies: the type of the theta atom, the type of the iota atom, and the position of the iota atom relative to the theta atom;
 - wherein a non-bonding intra-molecular contact is defined as an instance where the following conditions are satisfied:
 - $s - R_w \leq t$, where s is the separation between the theta and iota atoms, R_w is the sum of the van de Waals radii of the theta and iota atoms, and t is a predetermined threshold distance of typically 2.5 angstroms and preferably 0.8 angstroms; and
 - wherein in the case of proteins, the theta and iota atoms are on amino acid residues separated from each other by at least four residues on a linear polypeptide or are on separate polypeptide chains.
39. A method according to claim 38, wherein the method comprises sub-dividing the database to form groups of identified contacts in which the theta atom is one and only one of the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins and the iota atom is in one and only one of a plurality of non-overlapping groups obtained by sorting the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins into groups based on chemical similarity.
40. A method of generating a database for use in a method for designing a ligand *ab*

initio that will bind to a binding site of a macromolecular target, or of identifying a modification to a ligand for improving the affinity of the ligand to a binding site of a macromolecular target, comprising:

analysing the relative positions of atoms in each of a plurality of proteins or other biological macromolecules in order to identify instances of a non-bonding intra-molecular contact between a first atom referred to as a theta atom, and a second atom, referred to as an iota atom, of the protein or macromolecule; and

generating a database that for each identified contact specifies: the type of the theta atom, the type of the iota atom, and the position of the iota atom relative to the theta atom;

wherein a non-bonding intra-molecular contact is defined as an instance where the following condition is satisfied:

$s - R_w \leq t$, where s is the separation between the theta and iota atoms, R_w is the sum of the van de Waals radii of the theta and iota atoms, and t is a predetermined threshold distance of typically 2.5 angstroms and preferably 0.8 angstroms; and

wherein the method comprises sub-dividing the database to form groups of identified contacts in which the theta atom is one and only one of the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins and the iota atom is in one and only one of a plurality of non-overlapping groups obtained by sorting the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins into groups based on chemical similarity.

41. A method according to claim 39 or 40, wherein:

for each contact, the position of the iota atom is defined by geometrical reference to the position of the theta atom and to the positions of third and fourth atoms, the third atom being covalently bonded to the theta atom and the fourth atom being covalently bonded to the third atom, the method further comprising:

normalizing the coordinates of the iota atom, theta atom, third atom, and fourth atom of each contact as a group to generate a normalized coordinate group;

for each of one or more of the theta atom types, using the normalized coordinate groups for a plurality of contacts involving the theta atom type and a given iota atom type to generate a two-dimensional polar plot that represents a distribution of directions of the given iota atom, in terms of latitude and longitude, relative to the theta atom;

repeating the above step for different iota atom types;

comparing the resultant two-dimensional polar plots to identify groups of atom types that yield similar distributions of directions and using those groups as the groups based on chemical similarity to sort the 167 non-hydrogen atoms present in the 20 natural amino acids of proteins into the plurality of non-overlapping groups.

42. A method of generating a database according to any of claims 38-41, comprising:
extracting contact information from at least 2000 proteins or other biological macromolecules, the extracted information containing information about at least two million contact atom pairs.
43. A method of generating a database according to any of claims 38-42, comprising:
extracting contact information from at least 10000 proteins or other biological macromolecules, the extracted contact information containing information about at least ten million contact atom pairs.
44. A computer readable medium storing a database generated according to any of the previous claims.
45. A method according to any of claims 1 to 30 or using databases compiled according to any of claims 38 to 43, wherein: for a given antibody – antigen complex, specific mutations to amino acid residues in or around the antibody binding site are predicted to produce higher binding affinity of the antibody to the antigen.