nature structural & molecular biology

TECHNICAL REPORT

https://doi.org/10.1038/s41594-017-0019-z

Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas

Melis Kayikci[1,3*], A. J. Venkatakrishnan[1,4*], James Scott-Brown[1,5], Charles N. J. Ravarani[1], Tilman Flock[1,2,6] and M. Madan Babu[1*]

Visualizations of biomolecular structures empower us to gain insights into biological functions, generate testable hypotheses, and communicate biological concepts. Typical visualizations (such as ball and stick) primarily depict covalent bonds. In contrast, non-covalent contacts between atoms, which govern normal physiology, pathogenesis, and drug action, are seldom visualized. We present the Protein Contacts Atlas, an interactive resource of non-covalent contacts from over 100,000 PDB crystal structures. We developed multiple representations for visualization and analysis of non-covalent contacts at different scales of organization: atoms, residues, secondary structure, subunits, and entire complexes. The Protein Contacts Atlas enables researchers from different disciplines to investigate diverse questions in the framework of non-covalent contacts, including the interpretation of allostery, disease mutations and polymorphisms, by exploring individual subunits, interfaces, and protein–ligand contacts and by mapping external information. The Protein Contacts Atlas is available at http://www.mrc-lmb.cam.ac.uk/pca/ and also through PDBe.

Elucidating the structure of biomolecules has provided insights into how they carry out their function[1–3]. These insights have depended on advances in both methods for determining structures (initially X-ray crystallography[4] and NMR spectroscopy[5] and more recently electron microscopy[6]) and approaches for visualizing these structures. Historically, graphical representations of biomolecules have focused on covalent bonds that connect individual atoms, as in the ball-and-stick representation[7]. Such a representation emphasizes the 3D arrangement of atoms and the covalent bonds between them. This has been critical for understanding how function is mediated by the precise spatial localization of atoms in a biomolecule. Computational analyses of covalent bonds have been instrumental in uncovering the principles of protein architecture[8–11]. Similarly, the calculation of dihedral angles around the covalent bond, and their representation in the Ramachandran plot[12], charted the conformational landscape of polypeptides. The ribbon diagram, which was first popularized as the Richardson diagram[13], focused on the covalent backbone architecture and was revolutionary in providing a simplified protein structure representation. This enabled the identification of structural domains, establishment of the structure–function relationship, and classification of protein families[14–16]. In this manner, each of these representations centered on the covalent bond emphasizes a key aspect of structure and has formed a basis for deriving new insights and discoveries.

In addition to covalent bonds, however, non-covalent contacts between atoms of residues (residue contacts) are important for cooperative folding of biomolecules, for stability and conformational flexibility, and in molecular recognition. Representation of non-covalent contacts dates back to the 1970s in the form of contact matrices[17] and networks of contacts between amino acids in proteins[18]. More recently, a number of web tools that compute contact networks have facilitated progress in this area of research[19–23]. Network representations of non-covalent contacts, and their comparison among related structures, have provided insights into allosteric mechanisms, protein stability, conformational switching, ligand binding, and the determinants of protein fold and protein complex assembly[17,24–34]. Recently, we employed this approach to provide detailed molecular insights into the family of G-protein-coupled receptors (GPCRs) and G proteins[35–39]. In this manner, a residue-contact-based representation and analysis of protein structures enable us to identify critical contacts and holds the potential for understanding how biomolecules function in new ways and engineer their activity[40–42].

Here, we present the Protein Contacts Atlas (http://www.mrc-lmb.cam.ac.uk/pca/), a resource of non-covalent contacts from over 110,000 publicly available structures in the Protein Data Bank[43]. The goal of this atlas is to go beyond computing contact networks; for the exploration of contacts in structures, we developed interactive representations tailored for different scales of structural organization: atoms, residues, secondary structure, subunits, interfaces, and entire biological complexes (Fig. 1). The Protein Contacts Atlas also enables investigation of contacts within a single protein or a protein complex, or between a protein and nucleic acids, ligands, or other small molecules (Fig. 1). It also permits quantitative analyses of the residue-centric properties derived from the contact network along with externally obtained properties such as evolutionary conservation, thermostability measurements, etc. Outlier residues from the analyses that have the potential to inform follow-up experiments are compiled in a downloadable report. Here, we describe the visualization and analysis of non-covalent contacts in the Protein Contacts Atlas, which can be readily applied to any system, by focusing on diverse proteins involved in the GPCR signaling pathway.

Results

Computing non-covalent contacts. We identified non-covalent contacts by calculating the distance between each pair of atoms

[1]MRC Laboratory of Molecular Biology, Cambridge, UK. [2]Fitzwilliam College, University of Cambridge, Cambridge, UK. Present address: [3]Genomics England, London, UK. [4]Department of Molecular and Cellular Physiology, Department of Computer Science, and Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. [5]University of Oxford, Oxford, UK. [6]Paul Scherrer Institute, Villigen, Switzerland. Melis Kayikci and A.J. Venkatakrishnan contributed equally to this work. *e-mail: kayikci.melis@gmail.com; ajvenkat@stanford.edu; madanm@mrc-lmb.cam.ac.uk

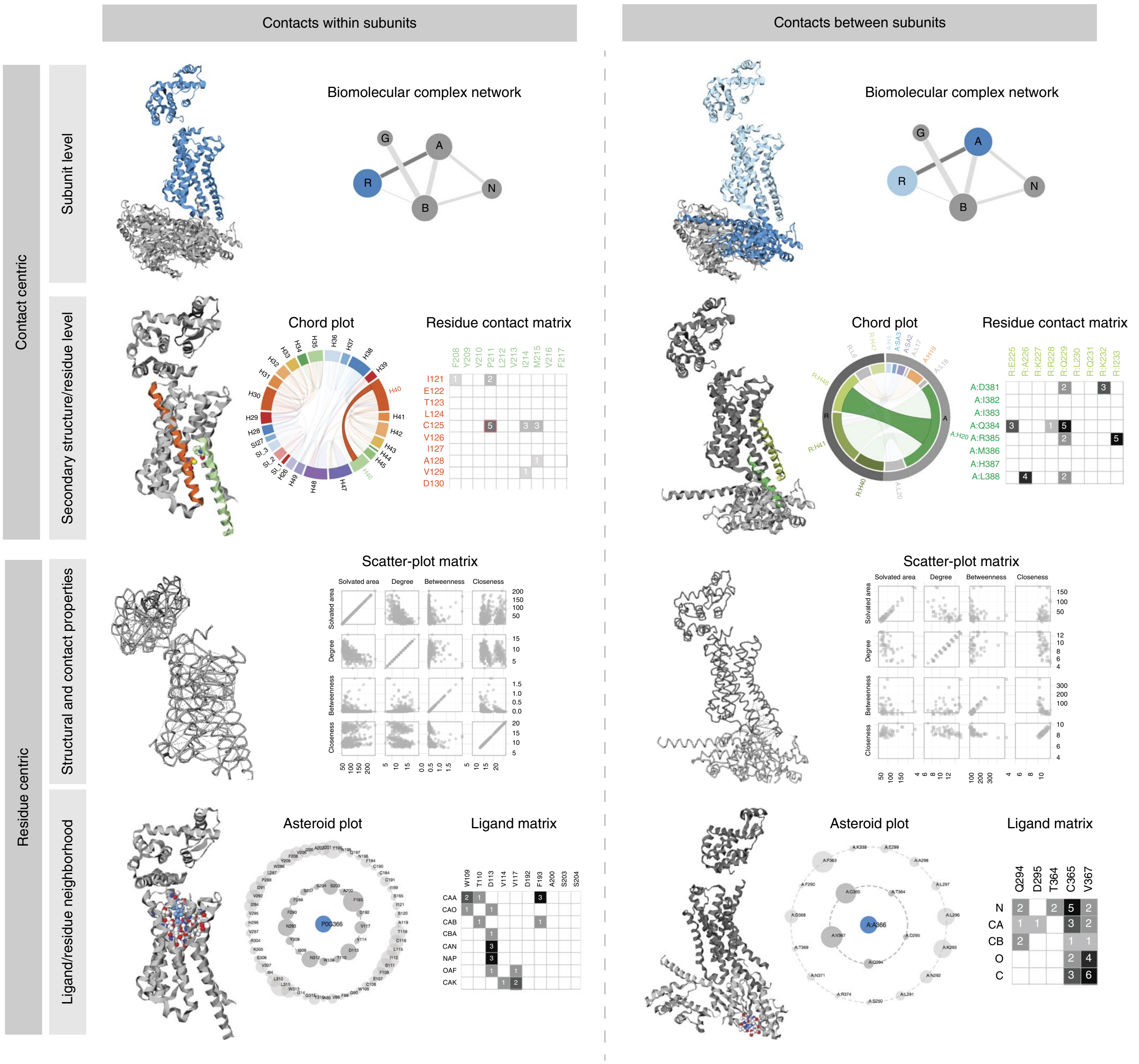


Fig. 1 | Visualization modes in the Protein Contacts Atlas. Summary of the representations for different scales of organization: atoms, residues, secondary structure, subunits, interfaces, and entire biological complexes.

using their atomic coordinates. We then compared the distances to the van der Waals radii of the corresponding atoms as determined by Chothia et al.[44]. The sum of the two atomic radii was subtracted from their distance and a contact was assigned if the resulting difference was less than a threshold. The set of all non-covalent atomic contacts defines a residue contact network, in which each node represents a residue, and a pair of nodes is joined by an edge if there is at least one non-covalent atomic contact between the corresponding residues (the number of such contacts is recorded as the edge weight). For each residue, we computed the local and global network centrality properties such as degree, closeness, and betweenness centrality (see Methods for details). We also quantified the solvent-accessible surface area (ASA; Å^2) for each residue based on the entire PDB file. Through this approach we identified ~2 billion non-covalent atomic contacts in over 110,000 crystal structures from the PDB (see Supplementary Data Set 5 for non-covalent contact statistics for each PDB file). As a general trend, we find that the number of non-covalent contacts scales linearly with the size of the molecule—i.e., the number of atoms and residues (Supplementary Fig. 1)—suggesting that this relationship can be used to infer the tightness of protein packing.

We provide different filtering and contact definition options whereby users can select a threshold, either in terms of absolute number of atomic contacts between residues and/or normalized with respect to the size of the amino acid, in order to define, view, and analyze stronger or weaker interactions. They can also filter contacts based on whether the atoms are from the main chains or side chains of residues and identify contacts by their type (i.e., hydrogen bonds, water-mediated hydrogen bonds, weak hydrogen bonds, ligand and metal complex interactions, salt bridges, hydrophobic interactions,

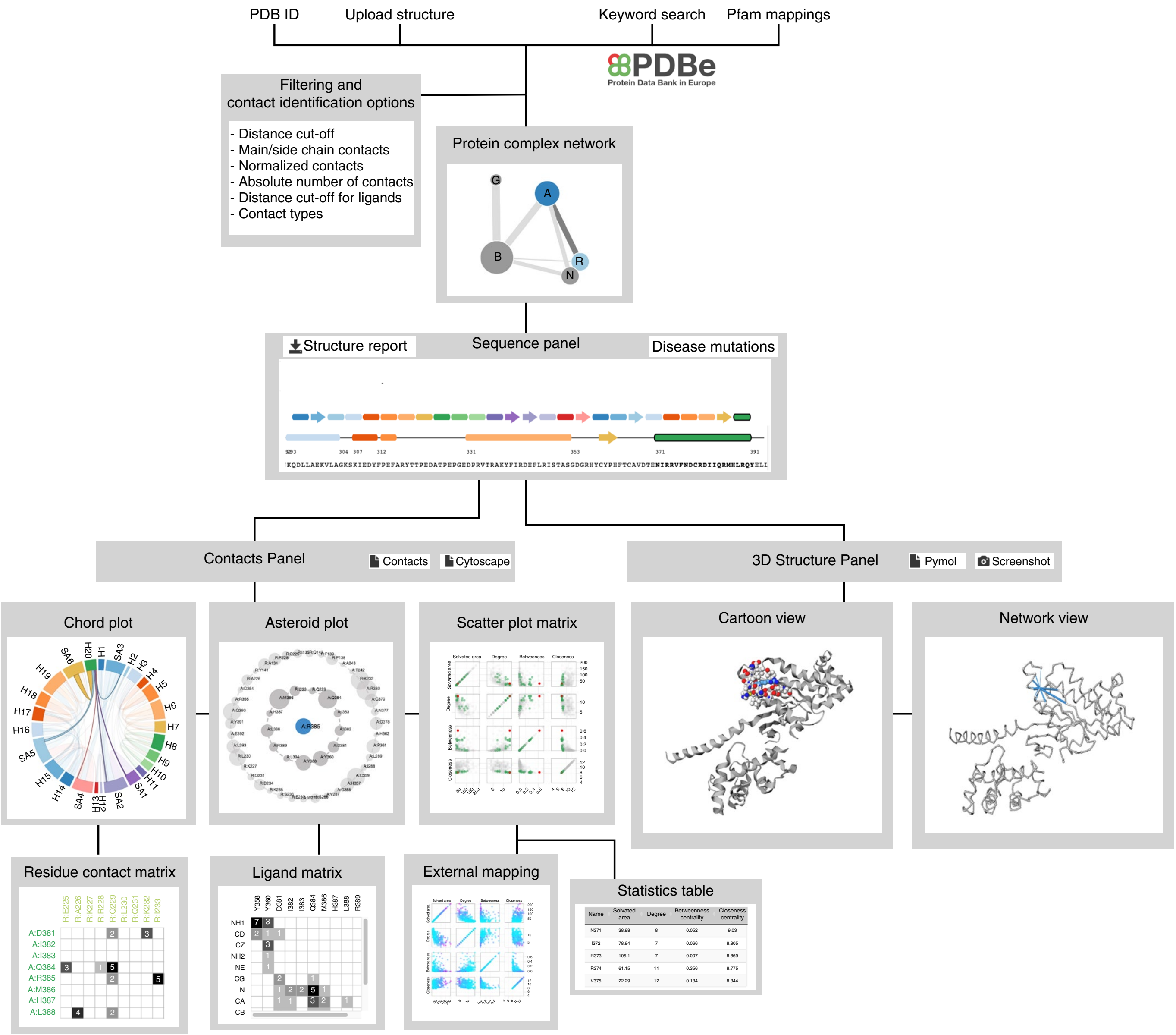


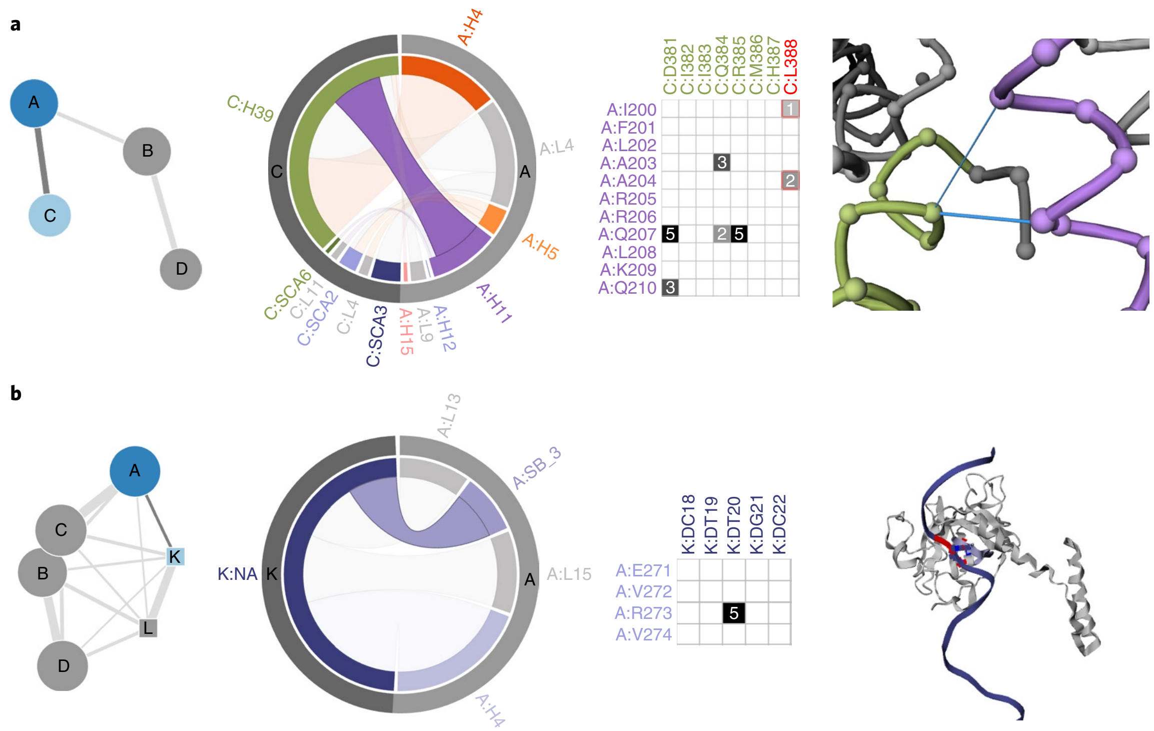
Fig. 2 | Protein Contacts Atlas framework. Summary of the Protein Contacts Atlas's framework for visualizing and analyzing contacts.

cation–pi interactions, pi–pi interactions, and other non-canonical contacts) using standard geometric considerations by employing the definitions used in Arpeggio[21]. Thus, the user has a number of different options to parameterize, filter, and/or choose specific contact types for visualization and analysis in the Protein Contacts Atlas.

Visualization of non-covalent contacts. We describe representations that allow intuitive and interactive visualization and analysis of residue contacts (Supplementary Note 1; Fig. 2; Supplementary Fig. 2). We highlight how representations of residue contacts of the same molecule at different scales of organization can provide new insights into structure and function that are not obvious from standard representations.

Biomolecular complex network enables visualization of contacts at the subunit level. How protein (or nucleic acid) subunits interact with each other in a complex is important for understanding the evolution and assembly pathways of the complexes[32,33]. The biomolecular complex network representation captures the interactions between subunits of a complex. In this network, the nodes denote individual subunits, which could be proteins or nucleic acids. The links between the nodes denote interaction interfaces between subunits (chains). The size of a node is proportional to the number of residues in the chain, and the thickness of the link is proportional to the number of residue contacts between the subunits. Such a simplified interactive representation of the entire complex provides an intuitive way to navigate and identify subunits or interfaces of interest for further investigation, particularly when investigating large, multi-subunit complexes (such as proteasome). Choosing a subunit, or an interface, takes the user to the "Visualization and Analysis" page (Supplementary Fig. 2).

Chord plot enables visualization of contacts at the secondary-structure level. The pattern of contacts between the different secondary-structure elements is key to determining tertiary structure, and hence protein function. 'Chord plots' depict all non-covalent contacts at the level of secondary structures. In a chord plot, every secondary structure (including the loops) is represented as an arc (nodes) in a



	C:D381	C:I382	C:I383	C:Q384	C:R385	C:M386	C:H387	C:L388
A:I200								1
A:F201								
A:L202								
A:A203				3				
A:A204								2
A:R205								
A:R206								
A:Q207	5			2	5			
A:L208								
A:K209								
A:Q210	3							

	K:DC18	K:DT19	K:DT20	K:DG21	K:DC22
A:E271					
A:V272					
A:R273			5		
A:V274					

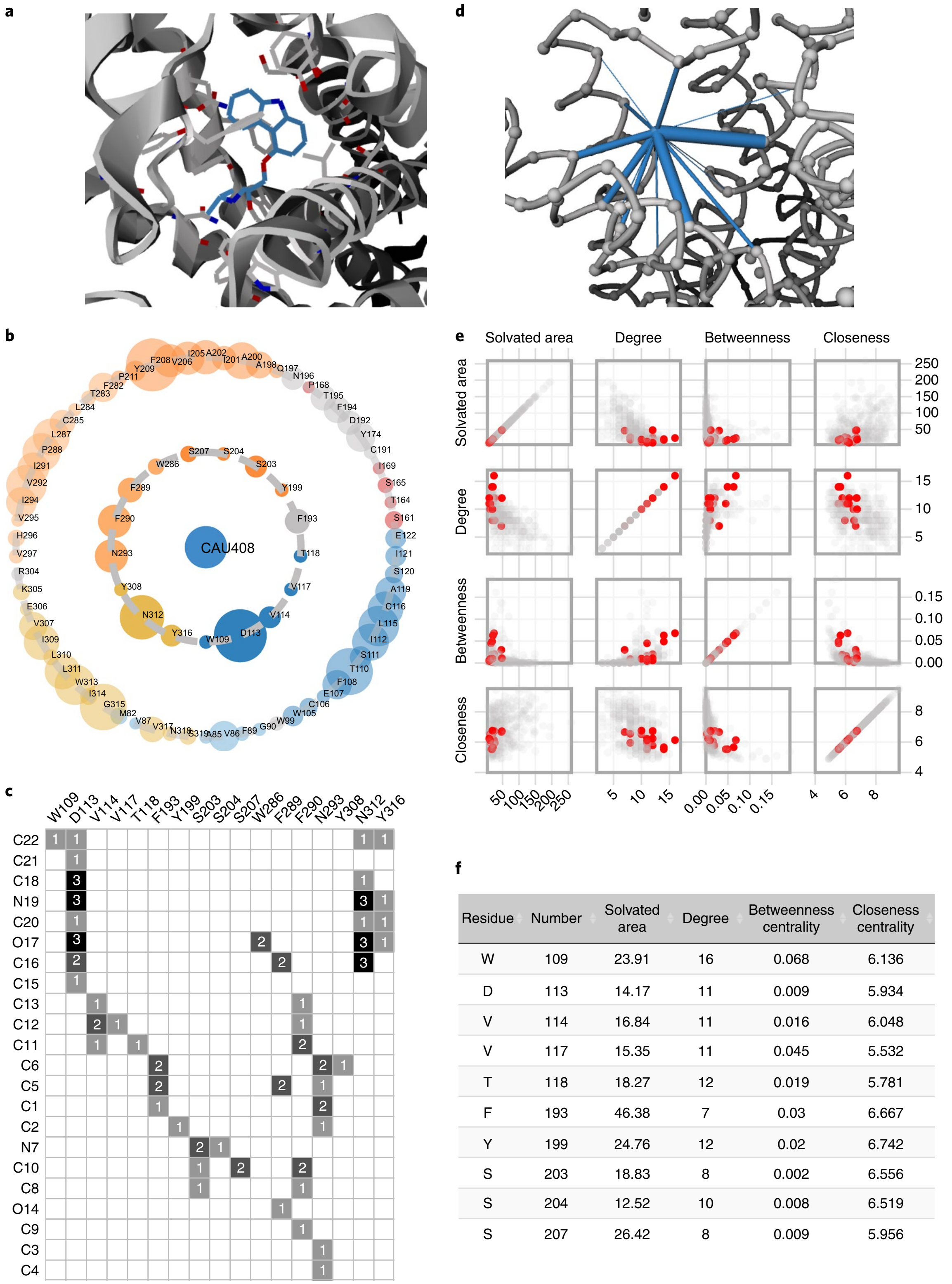
Fig. 3 | Visualization of protein–protein and protein–DNA interaction interfaces. a, The biomolecular complex network of the adenosine A2a–mini Gαs structure (PDB 5G53), with four chains as nodes and interactions between them as edges, is seen with chains A (A2a adenosine receptor; dark blue) and C (mini Gαs protein; light blue) highlighted. The chord plot of contacts between chains A (outer arc in light gray) and C (outer arc in dark gray) of the complex is seen next to it. The inner arcs show the secondary structures in their respective colors, with loops as light gray. The selected chord shows the contacts between helix 39 of the G protein (green) and helix 11 of the receptor (purple). The residue contact matrix of the interface is also shown, along with a network view of the receptor–G protein interaction interface (right). Positions that are mutated in pseudo-hypoparathyroidism (L388$^{G.H5.20}$; superscript denotes common G protein numbering system[36]) are shown in the network view and highlighted. Contacts are represented as blue edges and nodes are represented as spheres (using the Cα atom coordinates of the residues). **b**, The biomolecular complex network of p53 in complex with DNA (PDB 4MZR) with chain A (p53; circle, dark blue) and chain K (DNA; square, light blue) is highlighted. The selected chord highlights the contacts between sheet B3 of p53 (chain A; light gray, outer arc) and DNA (chain K; dark gray, outer arc). The residue contact matrix shows that there are five atomic contacts between Arg273 of p53 and T20 of the DNA strand. The 3D structure view of the protein–DNA complex with position Arg273 (red) that forms a part of the interaction interface and whose mutation is implicated in cancer is highlighted at right.

circular layout, and the contacts between the secondary structures are represented as chords (edges). The size of the arc is proportional to the number of residues within the secondary structure, and the thickness of the chord is proportional to the number of atomic contacts between them. The chord-plot representation provides information about the packing of the different secondary structures and helps identify the secondary structures that are highly connected in the protein structure.

Residue contact matrix enables visualization of contacts at the residue level. Identifying specific contacts between amino acids present on different secondary-structure elements helps with inferring the key residues that contribute to protein fold and function. The residue contact matrix presents the non-covalent contacts between residues in the secondary-structure elements (selected from chord plots) and displays the number of atomic contacts between them (Supplementary Fig. 2). Every cell in the matrix has a background color based on the number of atomic contacts. This allows easy identification of residue pairs that make a large number of contacts. The reside contact matrix is particularly useful to investigate the atomic details of interaction interfaces.

The multi-level visualization of non-covalent contacts in the context of a protein–protein interaction (between the adenosine A2a receptor and an engineered mini Gαs protein[45]) and a protein–nucleic acid interaction (between p53 and DNA[46]) is highlighted in Fig. 3. Such representations can also provide a non-covalent-contact-based context to generate testable hypotheses for understanding the molecular mechanisms of disease-associated mutations, as in pseudo-hyperthyroidism and Albright hereditary osteodystrophy (Supplementary Note 2).

Visualization and analysis of residue-centric contacts and properties. *Asteroid plot enables visualization of local neighborhood of residues and ligands.* Understanding the local neighborhood of ligands and residues in a structure can aid protein engineering, structure-based drug design, and interpretation of the effect of mutations (for example, via schematic diagrams of protein–ligand interactions generated by programs such as LigPlot[47]). 'Asteroid plots' provide an interactive representation of the atomic neighborhood of a selected ligand or a residue. The ligand or residue of interest is shown in the center as a node. All immediate residues that form a non-covalent contact (first-shell residues) are arranged in a circle around the ligand. The neighbors of each of the contacting residues that do not directly contact the ligand (second-shell residues) are arranged in a larger concentric circle. The sizes of the nodes in the inner and outer concentric circles denotes the number of atomic contacts. Clicking on any residue makes it the central node, and the asteroid plot for that entity is dynamically generated. This representation allows the identification of key residues that contact a ligand of interest in a structure. While the atomic details and the nature of the contact (for



	W109	D113	V114	V117	T118	F193	Y199	S203	S204	S207	W286	F289	F290	N293	Y308	N312	Y316
C22	1	1														1	1
C21		1															
C18		3														1	
N19		3														3	1
C20		1														1	1
O17		3									2					3	1
C16		2										2				3	
C15		1															
C13			1										1				
C12			2	1									1				
C11			1		1								2				
C6						2								2	1		
C5						2						2		1			
C1						1								2			
C2							1							1			
N7								2	1								
C10								1		2			2				
C8								1					1				
O14												1					
C9													1				
C3														1			
C4														1			

Residue	Number	Solvated area	Degree	Betweenness centrality	Closeness centrality
W	109	23.91	16	0.068	6.136
D	113	14.17	11	0.009	5.934
V	114	16.84	11	0.016	6.048
V	117	15.35	11	0.045	5.532
T	118	18.27	12	0.019	5.781
F	193	46.38	7	0.03	6.667
Y	199	24.76	12	0.02	6.742
S	203	18.83	8	0.002	6.556
S	204	12.52	10	0.008	6.519
S	207	26.42	8	0.009	5.956

Fig. 4 | Visualization and analysis of protein–ligand contacts. a, Ligand-binding pocket of β2 adrenergic receptor bound to the ligand carazolol (CAU408; blue; PDB 2RH1). All the directly contacting residues are shown as gray sticks. **b**, Asteroid plot with the ligand highlighted in blue (central node). Directly contacting residues (first-shell residues) are shown in the inner circle, and the residues that contact these but not the ligand (second-shell residues) are shown in the outer circle. The residues are colored according to their secondary structures, and the size of each circle is scaled to denote the number of atomic contacts. **c**, The ligand residue matrix shows the atoms (atom numbers are obtained from the PDB file) of the ligand as rows and the residues contacting the ligand as columns. Number of atomic contacts is also shown in the matrix. **d**, Ligand contacts are shown in the network view. **e**, All the ligand-contacting residues are highlighted in the scatter-plot matrix. **f**, Statistics table showing solvated area, degree, betweenness, and closeness centrality measures (the first 10 of 17 residues are shown).

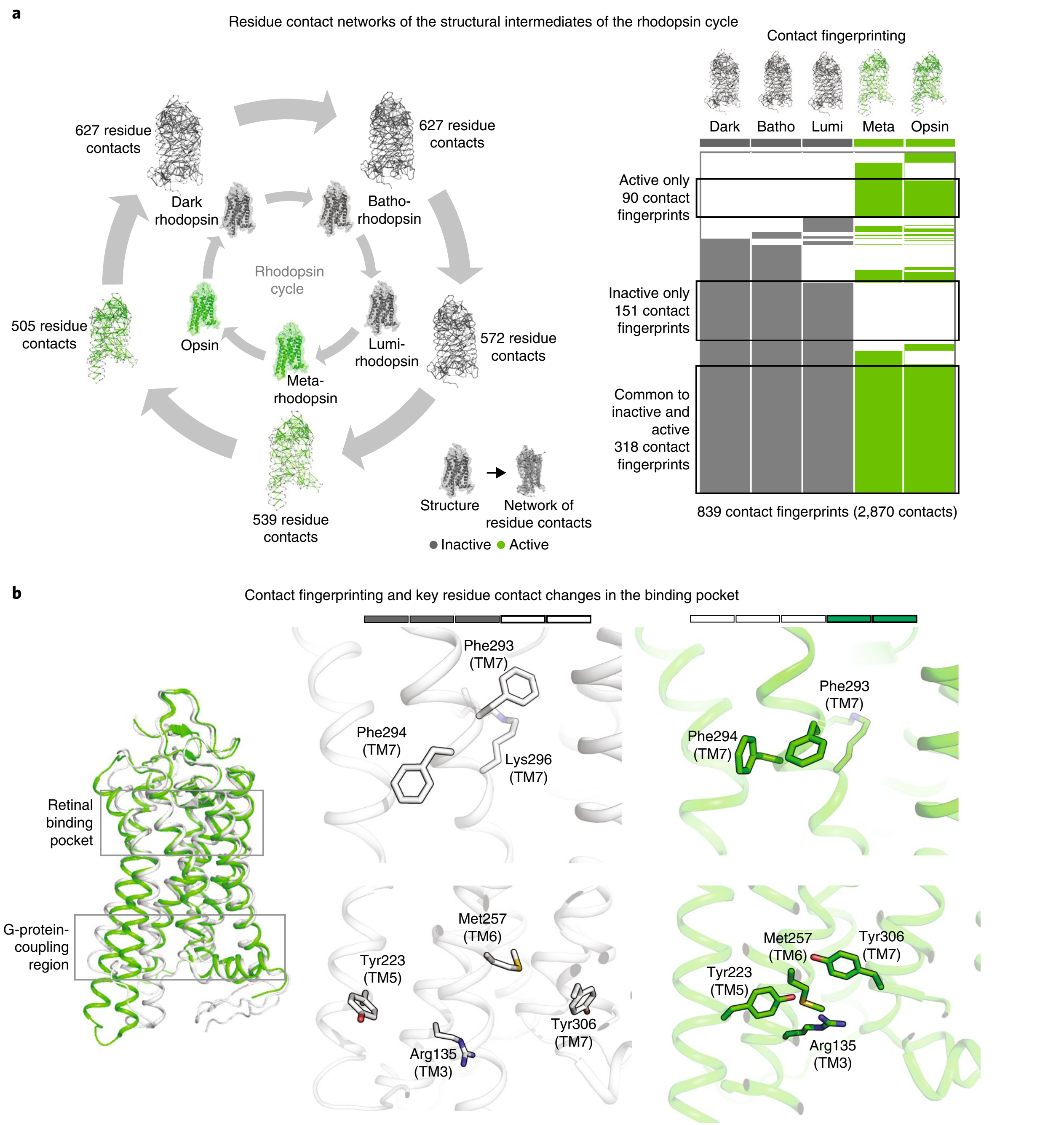
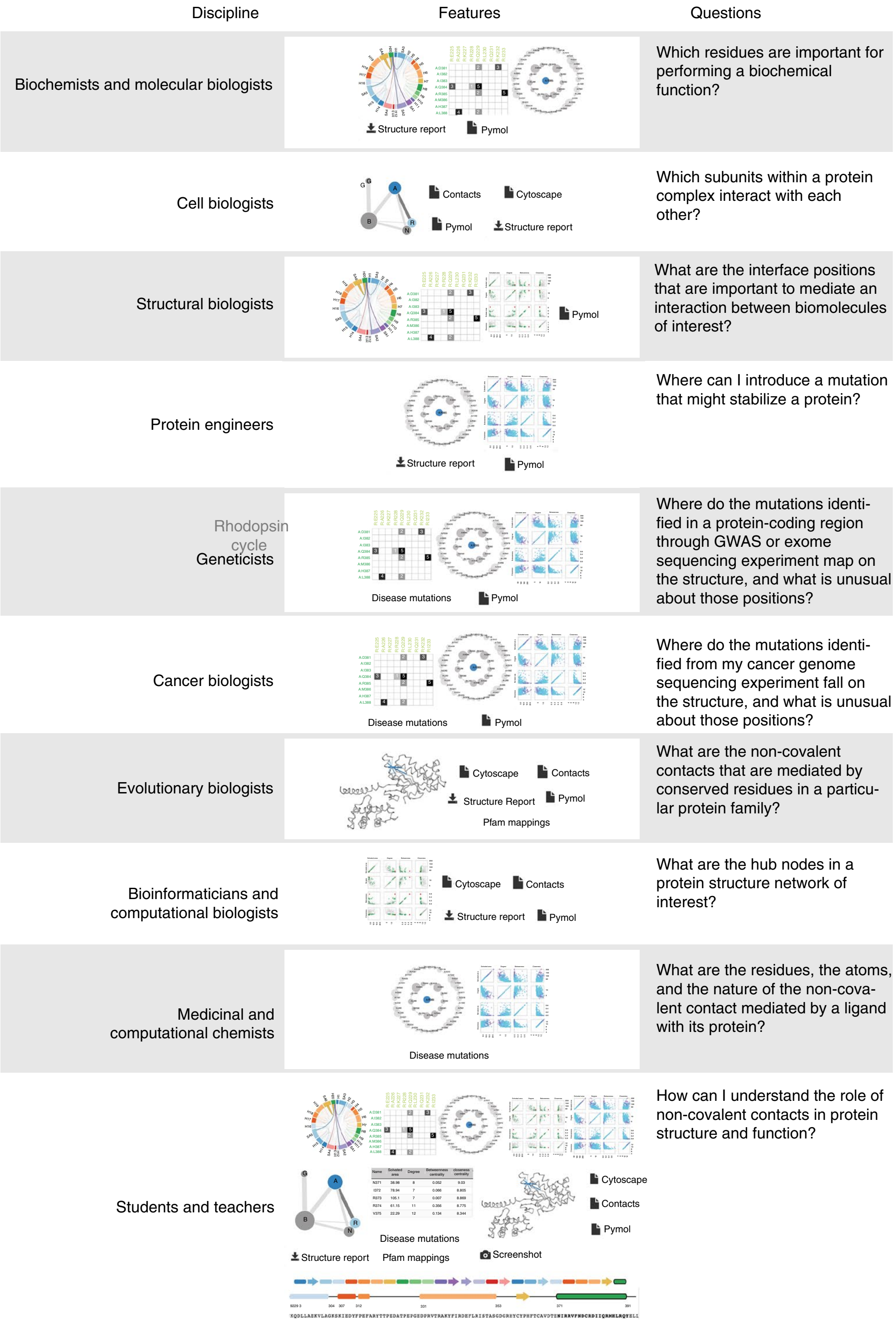


Fig. 5 | Comparison of the residue contact networks in the rhodopsin cycle. a, Residue contact networks of the structural intermediates of the rhodopsin cycle. In the residue contact networks, amino acid residues are denoted as nodes and the presence of contacts between pairs of residues is denoted as edges. 'Contact fingerprinting' and key residue contact changes in the binding pocket are displayed. For every residue contact within the ensemble, the presence or absence of an equivalent residue contact between equivalent positions across the rest of the states is recorded. This information is represented as an array of filled (contact present) and empty (contact absent) cells, which are referred to here as 'contact fingerprints'. **b**, Inactive-state-only contacts primarily map to the retinal-binding pocket and the G-protein-coupling region. Active-state-only contacts largely map to the region linking the two regions. A key structural change in the binding pocket involving aromatic contacts between Phe293 and Phe294 is shown in stick representation. In both panels, inactive states are shown in gray and active states in green.

example, main chain, side chain, etc.) are not shown in the asteroid plot, the contacting residues are highlighted in the 3D structure panel (Supplementary Note 1; Supplementary Fig. 2), enabling analysis of the nature of the contact. Furthermore, detailed information about the individual non-covalent atomic contacts between the individual atoms of the ligand and the contacting residues can be visualized as a ligand–residue interaction matrix from this subpanel (Fig. 1). Using the asteroid plot, we illustrate how the beta-blocker carazolol acts at its target human β2 adrenergic receptor (Fig. 4a–d; PDB 2RH1)[48]. Please see Supplementary Note 2 for a discussion on how such plots can provide a context for generating hypotheses of the molecular basis of a receptor polymorphism linked to asthma.

Scatter-plot matrix allows quantitative analysis of per-residue properties. Residues with distinct structural properties are important for function or are attractive sites for engineering. Quantifying different



Discipline	Features	Questions
Biochemists and molecular biologists	Structure report Pymol	Which residues are important for performing a biochemical function?
Cell biologists	Contacts Cytoscape Pymol Structure report	Which subunits within a protein complex interact with each other?
Structural biologists	Pymol	What are the interface positions that are important to mediate an interaction between biomolecules of interest?
Protein engineers	Structure report Pymol	Where can I introduce a mutation that might stabilize a protein?
Rhodopsin cycle Geneticists	Disease mutations Pymol	Where do the mutations identified in a protein-coding region through GWAS or exome sequencing experiment map on the structure, and what is unusual about those positions?
Cancer biologists	Disease mutations Pymol	Where do the mutations identified from my cancer genome sequencing experiment fall on the structure, and what is unusual about those positions?
Evolutionary biologists	Cytoscape Contacts Structure Report Pymol Pfam mappings	What are the non-covalent contacts that are mediated by conserved residues in a particular protein family?
Bioinformaticians and computational biologists	Cytoscape Contacts Structure report Pymol	What are the hub nodes in a protein structure network of interest?
Medicinal and computational chemists	Disease mutations	What are the residues, the atoms, and the nature of the non-covalent contact mediated by a ligand with its protein?
Students and teachers	Cytoscape Contacts Pymol Disease mutations Structure report Pfam mappings Screenshot	How can I understand the role of non-covalent contacts in protein structure and function?

Fig. 6 | How researchers from different scientific disciplines can make use of the Protein Contacts Atlas.

structural (such as surface area) and contact properties (such as number of contacts) on a per-residue basis, and analyzing their correlations, provides a way to identify outlier residues that might be important for structure and/or function (for example, a buried residue with a large number of contacts can be critical for protein stability)[26,28,49–51]. Per-residue external information such as sequence conservation, thermostability, disease mutations, etc., as well as the computed properties, can also be mapped onto the contact information and 3D structure for further analysis. Scatter plots display values for two variables for every residue in the chain(s) of interest: each residue is represented by a point, with the values of the variables determining its *x* and *y* coordinates. A matrix of scatter plots represents more than two variables using multiple scatter plots arranged in a grid, with one row and column per variable. The calculated properties, such as the ASA of the complex, network centrality measures (closeness and betweenness), and the degree, for each residue are plotted against each other in the scatter-plot matrix (Fig. 4e; see also Supplementary Note 2 and Supplementary Fig. 3 for highlighting multiple positions with disease mutations in rhodopsin onto the scatter plot).

Analysis of per-residue properties through an interactive statistics table. The interactive statistics table allows sorting the individual residues by any property (for example, residue name, number, ASA, degree, etc.), enabling the identification of residues with extreme values for a property. The table can be filtered by typing a residue (or multiple independent residues) in the text box, by selecting a bunch of data points directly in the scatter plot, or by clicking on a residue or secondary structure in the sequence panel (Supplementary Note 1; Fig. 4f). Clicking on any row will update the 3D structure panel with the selected residue, and the data can be downloaded in different formats (see sample PDF file in Supplementary Data Set 1).

Mapping external information for detailed per-residue analysis of structures. Combining the computed properties of individual residues with external information can help in identifying and characterizing functionally and structurally important regions or segments in protein structures. The Protein Contacts Atlas allows the importing and mapping of any external information that is relevant to a particular research question (for example, evolutionary conservation, disease mutations, thermostability, *b*-factors, post-translational modification sites, etc.). It automatically generates a template file that a user can download, complete, and upload to the website. The uploaded values are integrated with the statistics table, mapped onto all the relevant panels (including scatter-plot matrix, asteroid plot, and 3D structure panel), and color coded (from cyan, low, to magenta, high). In this manner, the Protein Contacts Atlas allows the user to integrate external and independently derived (i.e., orthogonal) information to make relevant inferences about a biomolecule of interest (see also Supplementary Note 2 and Supplementary Fig. 4 for interpreting stability measurements of point mutations in a G protein using this feature).

Structure report and downloadable data. A fully customizable report of the contact-based analysis of the selected chain of a structure can be downloaded (see example in Supplementary Data Set 3). It provides a summary of the session for a structure of interest, containing a screenshot of the current views of the structure from the 3D structure panel, the chord plot, an asteroid plot of the selected ligand, and the scatter-plot matrix from the contacts panel (Supplementary Note 1). Outlier residues are listed in tables that include those residues with the ten largest and ten smallest values of (i) ASA, (ii) degree, (iii) betweenness, (iv) closeness, and (v) number of atomic contacts (of the ligand, if there is one). The primary information about every non-covalent contact between atoms can be downloaded as a text file via the contacts panel (see example TXT file in Supplementary Data Set 2). The web resource can be queried in batch mode by retrieving structures based on their PFAM domain. The information can be downloaded in different formats including for stand-alone visualization with PyMOL (see Methods for details).

Rearrangement of residue contacts in rhodopsin cycle. To highlight how the analysis of residue contacts can be used to derive insights into protein function and mechanism, we present an analysis of the activation mechanism of rhodopsin. Overall, the analysis of the high-resolution structures of rhodopsin reveals a global rearrangement of non-covalent contacts underlying the first molecular events of vision.

Rhodopsin is a light-sensitive protein that is expressed in the eye and enables vision in dim light. In the absence of light, rhodopsin is bound to *cis*-retinal and is in an inactive state. Incidence of light catalyzes an isomerization reaction of retinal that leads rhodopsin to change shape to an activated form. This event triggers intracellular signaling cascades that ultimately culminate as an electrical impulse in the visual cortex of the brain. As non-covalent contacts are important for activation, investigating the organization of non-covalent contacts in rhodopsin is imperative for understanding how rhodopsin functions. During the activation process, rhodopsin forms a series of spectroscopically identifiable intermediate states, which when taken together constitute the rhodopsin cycle (Fig. 5a): dark rhodopsin, bathorhodopsin, lumirhodopsin, metarhodopsin (MI and MII), and free opsin. Extensive efforts in crystallography over the years have resulted in the determination of high-resolution structures of rhodopsin in these states. The availability of these structures provides an opportunity to systematically investigate the rearrangement of non-covalent contacts during rhodopsin activation.

Non-covalent contacts for the different states of the rhodopsin cycle were computed for bovine rhodopsin (PDB IDs 1U19, 2G87, 2HPY, 3PQR, 3CAP). The presence of contacts was compared across different states using contact fingerprinting (see Methods and Supplementary Data Set 4). A core network consisting of 318 residue contacts is present consistently in all the five states. This core network provides a state-independent platform for changes in non-covalent contacts in the rest of the protein. A separate network of 151 contacts connecting 163 residues is maintained exclusively in the inactive (dark, batho, and lumi) states. Upon the lumi-to-metarhodopsin transition, there is a major change in the organization of the contacts. The network of 151 contacts that was previously present in the inactive states is broken, and a new network of 90 contacts connecting 126 residues is formed exclusively in metarhodopsin and free opsin and is maintained until the end of the rhodopsin cycle.

In the dark, batho, and lumi states, the 151 contacts connecting 163 residues of the inactive states are largely localized near two regions in rhodopsin: (i) the retinal-binding pocket and the transmembrane–extracellular interface region and (ii) the region connecting the retinal-binding pocket and G-protein-binding site. In contrast, the 90 contacts of the active states are localized largely in the transmembrane region connecting the retinal binding pocket and G-protein-binding pocket. In the inactive states, in the retinal-binding pocket, one of the key contacts observed is between the aromatic ring of Phe293$^{7\times39}$ (TM7) and Lys296$^{7\times42}$ (TM7) linked to retinal (superscripts denote GPCRdb numbering[52,53]). Upon activation, this contact is broken and Phe293 engages in a contact with its adjacent amino acid Phe294$^{7\times40}$ (Fig. 5b). The change in the Phe293 side-chain orientation creates an opening between TM1 and TM7, and this local region has been associated with the channel that could be involved in the entry and exit of retinal[54]. Thus, Phe293 in the inactive states appears to be stabilizing the ligand through a contact, whereas the same residue in the active states creates an opening that could enable retinal's lateral entry and exit. In the G-protein-coupling region, some of the important residues in rhodopsin are Arg135$^{3\times50}$ (TM3), Tyr223$^{5\times58}$ (TM5), and Tyr306$^{7\times53}$ (TM7). In the inactive state, they are distal from each other. In the active state, Met257$^{6\times40}$ (TM6) contacts all these three residues (Fig. 5b). The Met257Tyr mutant form of rhodopsin is constitutively active[55].

Discussion

Representations of biomolecular structures highlighting specific aspects such as covalent bonds, volume, and surface area have had a profound impact on our understanding of function and on the development of new drugs[56]. For instance, space-filling models, Voronoi diagrams, and surface representations emphasize volume and surface area, which formed the basis for the identification and investigation of cavities and channels, electrostatic potentials, and interaction interfaces. Such cavities have been exploited for structure-based drug design. Drawing inspiration from how visualization and analysis of structures based on different representations revolutionized structure- and shape-based understanding of biomolecules, we developed representations that enable analysis of non-covalent contacts in biomolecules. Presenting representations of atomic contacts interactively and at different levels of organization (atoms, residues, secondary structures, and chains) alongside the classical 3D structural representation that is more familiar to most biologists provides the opportunity to investigate biomolecules in new ways. Given the likely general interest of this form of representation of protein structures, we have integrated the Protein Contacts Atlas via an API through the Protein Data Bank in Europe (PDBe)[57] website. The Protein Contacts Atlas has a modular design, allowing new features to be added easily. Future releases will include the ability to analyze all structures (including NMR and molecular dynamics models) and directly compare contacts between different structures (for example, conformational changes upon ligand binding).

The Protein Contacts Atlas allows scientists from diverse disciplines, including structural biologists, biochemists, molecular biologists, protein engineers, cancer biologists, medicinal and computational chemists, bioinformaticians, and geneticists, to address diverse questions (Fig. 6). Some typical tasks include mapping mutations from cancer genome sequencing experiments and genome-wide association studies, investigating protein structures for rational protein engineering, understanding how individual residues in homologous proteins evolve across homologs, identifying positions for mutational studies aimed at interrogating the function of biomolecules, and analyzing structures to derive new biological insights. Finally, the Protein Contacts Atlas can also serve as an excellent tool for teachers and students to explore and understand biological molecules at different levels of organization. We anticipate that the Protein Contacts Atlas will be a useful scientific resource as well as a learning platform that can fuel future research in biomedical sciences.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41594-017-0019-z.

Received: 23 February 2017; Accepted: 11 December 2017;
Published online: 15 January 2018

References

1. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. Kendrew, J. C. et al. Structure of myoglobin: a three-dimensional Fourier synthesis at 2Å resolution. *Nature* **185**, 422–427 (1960).
3. Perutz, M. F. et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* **185**, 416–422 (1960).
4. Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159**, 995–1014 (2014).
5. Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **8**, 923–925 (2001).
6. Cheng, Y. Single-particle cryo-EM at crystallographic resolution. *Cell* **161**, 450–457 (2015).
7. Ollis, W. D. Models and molecules. *Proc. R. Inst. G. B.* **45**, 1–31 (1972).
8. Perutz, M. F. The hemoglobin molecule. *Sci. Am.* **211**, 64–76 (1964).
9. Baldwin, J. & Chothia, C. Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *J. Mol. Biol.* **129**, 175–220 (1979).
10. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **37**, 205–211 (1951).
11. Richardson, J. S. β-Sheet topology and the relatedness of proteins. *Nature* **268**, 495–500 (1977).
12. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
13. Richardson, J. S. Early ribbon drawings of proteins. *Nat. Struct. Biol.* **7**, 624–625 (2000).
14. Levitt, M. & Chothia, C. Structural patterns in globular proteins. *Nature* **261**, 552–558 (1976).
15. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
16. Orengo, C. A. et al. CATH—a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
17. Nishikawa, K., Ooi, T., Isogai, Y. & Saitô, N. Tertiary structure of proteins. I. Representation and computation of the conformations. *J. Phys. Soc. Jpn.* **32**, 1331–1337 (1972).
18. Lesk, A. M. & Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270 (1980).
19. Chakrabarty, B. & Parekh, N. NAPS: Network Analysis of Protein Structures. *Nucl. Acids Res.* **44 W1**, W375–W382 (2016).
20. Seeber, M., Felline, A., Raimondi, F., Mariani, S. & Fanelli, F. WebPSN: a web server for high-throughput investigation of structural communication in biomacromolecules. *Bioinformatics* **31**, 779–781 (2015).
21. Jubb, H. C. et al. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* **429**, 365–371 (2017).
22. Doncheva, N. T., Assenov, Y., Domingues, F. S. & Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **7**, 670–685 (2012).
23. Piovesan, D., Minervini, G. & Tosatto, S. C. The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Res.* **44 W1**, W367–W374 (2016).
24. Vishveshwara, S., Brinda, K. V. & Kannan, N. Protein structure: insights from graph theory. *J. Theor. Comp. Chem.* **1**, 187–211 (2002).
25. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59–69 (2003).
26. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.* **2**, 0019 (2006).
27. Kornev, A. P., Haste, N. M., Taylor, S. S. & Eyck, L. F. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc. Natl. Acad. Sci. USA* **103**, 17783–17788 (2006).
28. Vishveshwara, S., Ghosh, A. & Hansia, P. Intra- and inter-molecular communications through protein structure network. *Curr. Protein Pept. Sci.* **10**, 146–160 (2009).
29. Fanelli, F., Felline, A. & Raimondi, F. Network analysis to uncover the structural communication in GPCRs. *Methods Cell. Biol.* **117**, 43–61 (2013).
30. Bhattacharyya, M., Ghosh, S. & Vishveshwara, S. Protein structure and function: looking through the network of side-chain interactions. *Curr. Protein Pept. Sci.* **17**, 4–25 (2016).
31. Fanelli, F., Felline, A., Raimondi, F. & Seeber, M. Structure network analysis to gain insights into GPCR function. *Biochem. Soc. Trans.* **44**, 613–618 (2016).
32. Ahnert, S. E., Marsh, A. J., Hernández, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).
33. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2**, e155 (2006).
34. Greene, L. H. & Higman, V. A. Uncovering network systems within protein structures. *J. Mol. Biol.* **334**, 781–791 (2003).
35. Venkatakrishnan, A. J. et al. Molecular signatures of G-protein-coupled receptors. *Nature* **494**, 185–194 (2013).
36. Flock, T. et al. Universal allosteric mechanism for Gα activation by GPCRs. *Nature* **524**, 173–179 (2015).
37. Venkatakrishnan, A. J. et al. Diverse activation pathways in class A GPCRs converge near the G-protein-coupling region. *Nature* **536**, 484–487 (2016).
38. Flock, T. et al. Selectivity determinants of GPCR-G-protein binding. *Nature* **545**, 317–322 (2017).
39. Hauser, A. S. et al Pharmacogenomics of GPCR drug targets. *Cell*, https://doi.org/10.1016/j.cell.2017.11.033 (2017).
40. Doncheva, N. T., Klein, K., Domingues, F. S. & Albrecht, M. Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci.* **36**, 179–182 (2011).
41. Martin, A. J. et al. RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics* **27**, 2003–2005 (2011).
42. Zhang, X., Perica, T. & Teichmann, S. A. Evolution of protein structures and interactions from the perspective of residue contact networks. *Curr. Opin. Struct. Biol.* **23**, 954–963 (2013).
43. Rose, P. W. et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucl. Acids Res.* **45**, D271–D281 (2017).
44. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253–266 (1999).
45. Carpenter, B., Nehmé, R., Warne, T., Leslie, A. G. & Tate, C. G. Structure of the adenosine A(2A) receptor bound to an engineered G protein. *Nature* **536**, 104–107 (2016).

46. Emamzadah, S., Tropia, L., Vincenti, I., Falquet, B. & Halazonetis, T. D. Reversal of the DNA-binding-induced loop L1 conformational switch in an engineered human p53 protein. *J. Mol. Biol.* **426**, 936–944 (2014).
47. Laskowski, R. A. & Swindells, M. B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* **51**, 2778–2786 (2011).
48. Cherezov, V. et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318**, 1258–1265 (2007).
49. Mendes, H. F., van der Spuy, J., Chapple, J. P. & Cheetham, M. E. Mechanisms of cell death in rhodopsin retinitis pigmentosa: implications for therapy. *Trends Mol. Med.* **11**, 177–185 (2005).
50. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci.* **15**, 2120–2128 (2006).
51. Soundararajan, V., Raman, R., Raguram, S., Sasisekharan, V. & Sasisekharan, R. Atomic interaction networks in the core of protein domains and their native folds. *PLoS ONE* **5**, e9391 (2010).
52. Isberg, V. et al. Generic GPCR residue numbers—aligning topology maps while minding the gaps. *Trends Pharmacol. Sci.* **36**, 22–31 (2015).
53. Isberg, V. et al. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **44 D1**, D356–D364 (2016).
54. Hildebrand, P. W. et al. A ligand channel through the G protein coupled receptor opsin. *PLoS ONE* **4**, e4382 (2009).
55. Deupi, X. et al. Stabilized G protein binding site in the structure of constitutively active metarhodopsin-II. *Proc. Natl. Acad. Sci. USA* **109**, 119–124 (2012).
56. O'Donoghue, S. I. et al. Visualizing biological data-now and in the future. *Nat. Methods* **7 (Suppl.)**, S2–S4 (2010).
57. Velankar, S. et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucl. Acids Res.* **44**, D385–D395 (2016).

Acknowledgements

We thank A. Lesk, C. Chothia, S. Balaji, H. Harbrecht, I. Huppertz, M. Ouédraogo, G. Chalancon, A. Morgunov, A. Murzin, A. Andreeva, G. de Baets, R. Peer, S. Chavali, A. Sente, N. S. Latysheva, A. Gunnarsson, and A. Hauser for their comments on this work. We thank T. Nakane for his inputs on structure visualization using WebGLMol. This work was supported by the Medical Research Council (MC_U105185859; M.M.B., M.K., C.N.J.R., T.F., A.J.V., and J.S.-B.), the LMB-Cambridge scholarship (A.J.V.), the St. John's College Benefactor scholarship (A.J.V.), the AFR scholarship from the Luxembourg National Research Fund (C.N.J.R.), and the Boehringer Ingelheim Fond (T.F.). T.F. is a Research Fellow of Fitzwilliam College, University of Cambridge, UK. M.M.B. is a Lister Institute Research Prize Fellow.

Author contributions

M.K. collected the data, developed the computational pipeline, and built the web server. A.J.V. designed the prototype of the representations with M.M.B. A.J.V., M.K., J.S.-B., and M.M.B. optimized the representations, and M.K. and J.S.-B. implemented the representations. M.K. and A.J.V. performed the GPCR analyses. J.S.-B. made the prototype of the web server. C.N.J.R. and T.F. helped with the web server and analyzing examples. M.K., C.N.J.R., and A.J.V. independently wrote separate drafts of the manuscript. M.K., A.J.V., and M.M.B. wrote the final manuscript with critical inputs from C.N.J.R., J.S.-B., and T.F.; M.K., A.J.V., and C.N.J.R. prepared the figures. A.J.V. and M.M.B. conceived and planned the project. M.K., A.J.V., and M.M.B. executed the project. M.M.B. supervised the project.

Competing interests

The authors declare no competing financial interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-017-0019-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.K. or A.J.V. or M.M.B.

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

Methods

Data preprocessing. New structures added to the PDB are preprocessed in a batch process that is run every 6 months. Structures uploaded by a user undergo the same preprocessing steps, which may take up to 5–7 min depending on the size of the file. First, secondary structures are determined using DSSP if this information is not available in the PDB file[58,59]. Non-covalent contacts between atoms are computed using a custom C++ program. Solvent-accessible surface area (ASA) is also calculated for each residue using an external program (POPS[60]). For contact identification, we calculated the distance between each pair of atoms based on the coordinates provided in the PDB file. The sum of the two atomic radii (defined by Chothia et al.[44]) was subtracted from this distance and a contact was assigned if the resulting difference was less than a threshold. By default, the threshold used to define if two atoms are contacting is 0.5 Å, but the user can choose any value in the range of 0 to 1 Å. In the case of ligands (including all non-amino acid residues except water (for example, ions)) or water molecules in the HETATM record in the coordinate file, a contact is assigned if the distance is <4 Å. The user has the option to provide different distance cutoff values for the ligands.

Filtering and contact identification options. The user has different options to view and analyze contacts:

1.A filtering step to select a threshold in terms of absolute number of atomic contacts between residues. In this option, the user selects the minimum number of atomic contacts between pairs of residues to view and analyze. After selecting this parameter, the filtering is done within the C++ program. The user has all the options as before for viewing and analyzing the contacts.

2.A filtering step to view and analyze contacts involving main chain–main chain, side chain–side chain or main chain–side chain atoms.

3.A filtering step to select contacts that are normalized with respect to the size of the amino acid. For the normalization, we used a previously published approach[19,61]. Briefly, we first identified all non-redundant crystal structures in the PDB (dated 19.06.17) from the NCBI database (ftp://ftp.ncbi.nih.gov/mmdb/nrtable/) using a resolution cut-off of 2 Å. This resulted in 48,856 structures (95,159 chains). We then calculated the average of the maximum number of atomic contacts made by each of the 20 amino acids in these structures (please see Supplementary Fig. 5). This was done using the precomputed results available as JSON files in Protein Contacts Atlas. Using this as our reference table in the C++ program, we computed normalized contacts for any structure provided by the user using the following formula:

$$\text{normalized weight} = (\text{number of side chain atomic contacts} \;/\; \text{sqrt}(\text{norm}_{\text{res1}} \times \text{norm}_{\text{res2}})) \times 100$$

where "number of side chain atomic contacts" is the number of atomic contacts between the side chains of two residues where the distance between two atoms is smaller than 4 Å and norm_res1 and norm_res2 are the values taken from the calculated table. After the normalized weight between the residues is calculated, any interaction which has the normalized weight smaller than the threshold chosen by the user is filtered.

A distribution of normalized weights from the 48,856 non-redundant structures is provided as a guide for users to choose the threshold (Supplementary Data Set 6). This distribution is also available in the "more info" section of the website while choosing this option for filtering contacts.

If the user changes the default threshold, it only affects filtering options 1, 2, and 3 above.

4.Protein Contacts Atlas can also calculate hydrogen bonds, water-mediated hydrogen bonds, weak hydrogen bonds, ligand and metal complex interactions, salt bridges, hydrophobic interactions, cation–pi interactions, pi–pi interactions, and other non-canonical contacts using standard geometric considerations by employing Arpeggio[21].

Statistical analysis. For each residue and heteroatom in the residue contact network, Protein Contacts Atlas calculates a range of network centrality measures[62] (betweenness centrality, closeness centrality and degree), which measures its "importance" in the overall contact network. Centrality measures were computed using the SNAP library written in C++ (see their documentation at https://snap.stanford.edu/snap/doc/snapuser-ref/index.html for detailed definitions). There are two ways of computing the network statistics: with and without water molecules. By default, the network statistics with water molecules are not shown; however, the user has the option to view them with water. Betweenness centrality defines how many times the residue of interest falls in the shortest paths connecting other residues. The betweenness centrality values are normalized using the following formula: normalized betweenness for each residue = betweenness for each residue × (2/(([total number of residues] − 1) × ([total number of residues] − 2))). This measure expresses the amount of "control" exerted by that residue over the contacts between other residues in the network[63,64]. Closeness centrality is defined as the inverse of the sum of distances of the residue of interest from all other residues[65,66] and is normalized within SNAP. In this case, the more buried a given residue is, the more contacts it has, and the closer it is to other residues. The closeness centrality is a measure of how quickly "information" spreads from a given residue to other residues in the network. Both measures show how central the residue is with regard to the whole residue contact network. The degree of a residue is the number of other residues it contacts.

Data visualization. Preprocessed results are stored in JSON files (http://json.org/) and are used to produce the interactive visualizations. JSON was chosen because it is a simple format that is easy to generate and parse. A typical JSON file includes the name, numbers and weights (number of atomic contacts) of the residues, secondary structure elements or loops to which the residues belong, whether it is a heteroatom (ligands, water) or not, and which chain they belong to. The file also contains the contacts separately for each residue. This includes the contacting residue pairs, the contacting atoms in each residue, the distances between the contacting atoms, the types of atoms (main or side chains) and the total number of atomic contacts within the residue pair. Finally, the file includes the secondary structure definition (start and end positions and names, for example, A:HELIX14, A:SHEETB_2, B:LOOP1). The JSON files are used as an input to visualize the contacts in the browser using JavaScript, HTML and CSS. The Bootstrap framework (http://getbootstrap.com/) is used for the overall page layout, and the D3.js library (http://d3js.org/) is used to produce interactive graphs and plots, including the chord plot, the asteroid plot and the scatter-plot matrix.

Calculations of residue contacts and contact fingerprinting for the Rhodopsin case study. The residue contacts were computed for the structures representing the different intermediates of the rhodopsin cycle. A residue contact between a pair of residues is defined as present when the distance between any two atoms from the residue pair is less than the sum of their van der Waals radii plus a cut-off distance of 0.5 Å[35–37]. We analyzed the presence of residue contacts between structurally equivalent residues across the different conformational states of rhodopsin. The functional importance of a given residue contact across conformational states can be estimated based on the extent to which structurally equivalent contacts are maintained consistently. For every residue contact within the ensemble, the presence or absence of an equivalent residue contact between structurally equivalent positions across the rest of the states was recorded. This information is stored as a bit string of 1 s (present) and 0 s (absent), which are referred to as "contact fingerprints"[37]. Identifying contact fingerprints that represent consistently maintained residue contacts across and between conformational states enabled us to identify the key rearrangements of residue contacts during rhodopsin activation.

Description of the visualization features. The PDB file itself may not reflect the biological unit. Therefore, a PDBe PISA link is also provided for that PDB file (for example, http://www.ebi.ac.uk/pdbe/pisa/cgi-bin/piserver?qa=3sn6). The link provides access to the PDB coordinates to the different plausible biological units of the proteins involved in the complexes. The user can then choose the relevant assembly of interest and upload the file to Protein Contacts Atlas for visualization and analysis. Upon selecting a structure, a page displays the entire 3D structure of the molecule in cartoon representation and a corresponding protein complex network. The main page has three interlinked panels, displaying representations of the sequence, the 3D structure, and the non-covalent contacts (see Supplementary Note 1 for details). The biomolecular complex network is always shown on the top left, providing the opportunity for the user to easily switch between chains or interfaces. For the chord plot, moving the cursor over a chord or arc on the contacts panel increases the transparency of the other secondary structures, making it easier to identify and investigate contacts between secondary structures that are far away in the protein sequence. If the user selects an interface, this view provides information about the secondary structures that interact between chains and the thickness is indicative of the strength of the interface. The colors used for secondary structure representation are consistent across different panels (Supplementary Fig. 2 and Supplementary Note 1). Users have the option to manually define "super secondary structure elements" and/or adjust the exact definition of a secondary structure. For the residue contact matrix, clicking on the individual elements within matrix (which is accessed by clicking on a chord first) highlights the relevant contacting residues in the 3D structure panel, providing the opportunity to investigate the chemical nature of the contact (for example, side-chain or main-chain contacts). The ligand contacts, which can be seen in the Ligands and Residues sub-panel can be independently visualized and analyzed by downloading a PyMOL script that is provided in the 3D structure panel.

Description of the analysis features. In the scatter plot matrix, the color spectrum for the different properties can also be set by the user to obtain publication quality images and/or to visualize the 3D structure for detailed analysis (by downloading the updated PyMOL session file). In the Per Residue Statistics sub-panel, clicking and dragging the cursor over a specific region of any scatter plot selects the data points in this region, and simultaneously highlights the same set of residues in the other scatter plots, the 3D structure panel and the sequence panel. Individual residues of interest are either selected by typing in the residue number in a text box or by clicking on a residue in the sequence panel. Multiple independent residues (for example, several disease mutations) can be selected by typing in the residue numbers separated by a comma in the text box. This highlights the residue(s) in red in all the scatter plots and in the 3D structure panel.

Download file formats and options for accessing contact information for several structures. The contacts file for the individual structures (with all their chains) contains information about the chain, secondary structure, residue name and number, number of atomic contacts for each residue pair, atom names, types (main chain or side chain atom) and the distance (Å) between the contacting atoms in text format. If a user is interested in a particular protein family but does not have a list of PDB codes of structures that contain the domain, they can use the "PDBs by PFAM" option (within "Advanced Options") to download contacts of all structures that contain a PFAM domain of interest using the default options for contact definition. In addition, users can also download this information in a simple interaction file (SIF) format, which serves as an input for Cytoscape[67], a popular open source software for complex network analysis. The user can also download high-resolution screenshots of images from the contacts panels in support vector graphics (SVG) format, screenshots of the 3D structure in portable network graphics (PNG) format, and PyMOL session files for stand-alone visualization in PyMOL.

Web server specifications. Protein Contacts Atlas has been tested on Chrome, Firefox and Safari (versions 6.0 and higher) and works best in Chrome.

Life Sciences Reporting Summary. A Life Sciences Reporting Summary for this article is available.

Code availability. Source code for the project is available at https://github.com/pandora2017/protein_contacts_atlas.

Data availability. Protein Contacts Atlas is available online at http://www.mrc-lmb.cam.ac.uk/pca/. Users can access the information for any PDB file directly from the link www.mrc-lmb.cam.ac.uk/pca/redirect/3sn6. A detailed tutorial can be accessed via the "Quick Tutorial" button. Other data are available upon request.

References

58. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
59. Touw, W. G. et al. A series of PDB-related databanks for everyday needs. *Nucl. Acids Res.* **43**, D364–D368 (2015).
60. Cavallo, L., Kleinjung, J. & Fraternali, F. POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucl. Acids Res.* **31**, 3364–3366 (2003).
61. Kannan, N. & Vishveshwara, S. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **292**, 441–464 (1999).
62. Costa, L. F., Rodrigues, F. A., Travieso, G. & Villas Boas, P. R. Characterization of complex networks: a survey of measurements. *Adv. Phys.* **56**, 167–242 (2007).
63. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977).
64. Yoon, J., Blumer, A. & Lee, K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics* **22**, 3106–3108 (2006).
65. Bavelas, A. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.* **22**, 725–730 (1950).
66. Sabidussi, G. The centrality of a graph. *Psychometrika* **31**, 581–603 (1966).
67. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

nature research

Corresponding author(s): Melis Kayikci, AJ Venkatakrishnan, M. Madan Babu

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

N/A

2. Data exclusions

Describe any data exclusions.

N/A

3. Replication

Describe whether the experimental findings were reliably reproduced.

N/A

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	□	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	□	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	□	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this

C++ code was written to compute atomic contacts for over 100,000 crystal study.

structures. Javascript, HTML, CSS was used to create the webserver. Python was used to calculate statistics for atomic and residue contacts between and within chains.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A