Coevolutionary data-based interaction networks approach highlighting key residues across protein families: The case of the G-protein coupled receptors

Computational and Structural Biotechnology Journal 18 (2020) 1153–1159

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Coevolutionary data-based interaction networks approach highlighting key residues across protein families: The case of the G-protein coupled receptors

Filippo Baldessari [a], Riccardo Capelli [b,*], Paolo Carloni [b], Alejandro Giorgetti [a,b]

[a] Department of Biotechnology, Università di Verona, Ca Vignal 1, strada Le Grazie 15, I-37134 Verona, Italy
[b] Computational Biomedicine Section, IAS-5/INM-9, Forschungzentrum Jülich, Wilhelm-Johnen-straße, D-52425 Jülich, Germany

ARTICLE INFO

Article history:
Received 27 February 2020
Received in revised form 1 May 2020
Accepted 6 May 2020
Available online 15 May 2020

Keywords:
GPCRs
Coevolution
Interaction network
Conformational states
Functionally relevant residues

ABSTRACT

We present an approach that, by integrating structural data with Direct Coupling Analysis, is able to pinpoint most of the interaction hotspots (i.e. key residues for the biological activity) across very sparse protein families in a single run. An application to the Class A G-protein coupled receptors (GPCRs), both in their active and inactive states, demonstrates the predictive power of our approach. The latter can be easily extended to any other kind of protein family, where it is expected to highlight most key sites involved in their functional activity.

1. Introduction

The discovery of correlated mutations in protein families across different organisms has shown to provide valuable information on the functional role of residues [1]. These mutations arise from evolutionary pressure, that drives the changes to enhance stability and/or biological function. Starting from the correlation between mutations in sequence alignments (*i.e.*, coevolutionary analysis, CA), one can infer that the destabilization induced by a single-point mutation can be attenuated or even counterbalanced by a corresponding mutation in a different portion of the sequence. Different CA approaches include DCA [2–4], plmDCA [5], GREMLIN [6], PSICOV [7], and many others [8,9]. These methods have been used for different goals [10]: from the definition of coarse-grained force fields for molecular simulation [11], to the prediction of mutation energetics [12,13], the direct inference of 3D structures [3], refinement of structure prediction [14,15], and the investigation of protein–protein interactions [16,17].

Here we introduce a structure-based CA that identifies in a single run highly structurally and/or functionally relevant residues across a protein family. This is achieved integrating structural information on a modified version of Lui and Tiana's [12] approach. The latter uses internal interaction networks to uncover frustrated interactions and mutation free energy differences for a specific protein. Our protocol has several advantages. Unlike previous approaches, mainly based on single protein domains (usually obtained from PFAM [18]), our protocol includes entire proteins in the calculations. In addition, it provides insight on different conformational states of proteins (unlike "classical" DCA analysis [2], which is based on pure sequence information), such as receptor active/inactive states or ion channel close/open states. Finally, and most importantly, it can be applied to large sparse families, i.e. with very large sequence variability due to evolution. Applications to the sparse and fairly well-structurally characterized [19] subfamily, "class A" of the human G-Protein Coupled Receptor (hGPCRs) superfamily, shows the predictive power of the approach: in a single run, it identifies coevolutionary related hotspots, previously pinpointed by techniques other than CA [20,21], integrating also structural information to highlight differences in distinct conformational states. These are fundamental structural/-functional residues or correlated with diseases. The protocol is totally general and can easily be extended to other subfamilies of GPCRs, from organisms other than *Homo Sapiens*, as well of other large receptor families with large intrinsic variability, like the pentameric ligand-gated ion channels (pLGICs) or the voltage-gated ion channels.

* Present Address: Department of Applied Science and Technology (DISAT), Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy.
E-mail addresses: r.capelli@fz-juelich.de, riccardo.capelli@polito.it (R. Capelli).

https://doi.org/10.1016/j.csbj.2020.05.003

2. Results

The approach

In the following sections, we briefly describe our multistep strategy (Fig. 1). Our approach uses structural information in two different stages of the protocol, i.e. during the multiple sequence alignment (MSA) generation and in the construction of the interaction matrix, in contrast to most of the previously used coevolutionary approaches, that usually do not consider this information.

Below we report details of each of these steps.

2.1. Experimental data

Let us consider the general case for which sequences across a given family are highly diverse. In this case several bottlenecks, starting from the MSA generation, can be found. Thus, the use of state of the art methodologies can be applied to overcome these difficulties. We consider here an alignment formed by M sequences composed by L residues σ_i (where i is the kind of amino acid) obtained form curated databases (Uniprot [22], Pfam [18], or GPCRdb [23]). The MSA can be generated using an algorithm, Promals3D [24], that utilizes structural data, and thus can alleviate the difficulty in aligning families that displays sparse sequences.

2.2. Interaction analysis – DCA and Sequence-specific interaction matrix

Our coevolutionary analysis protocol is based on Direct Coupling Analysis [2,3]. The basic assumption of such technique is that the interaction between different residues can be written as

$$U(\{\sigma_i\}) = \sum_{i<j} \mathscr{E}_{ij}(\sigma_a, \sigma_b) + \sum_i h_i(\sigma_a) \tag{1}$$

where σ_i is the amino acid at i-th position of the sequence, $\mathscr{E}_{ij}(\sigma_a, \sigma_b)$ is the two-bodies term (analogous to the interaction term of a Potts' model) that contains the interaction energy between residues σ_i and σ_j at position i and j of the alignment, respectively, and $h_i(\sigma_a)$ is a one-body potential that can act on the residues (analogous to the field term of a Potts' model).

The key quantity here is the two-bodies term $\mathscr{E}_{ij}(\sigma_a, \sigma_b)$, that contains the information needed to build the interaction graph, leading, eventually, to the identification of the hotspots (see below). The frequency counts of pairs $f_{ij}(\sigma_a, \sigma_b)$ and single $f_i(\sigma_a)$ residues of a sequence with n amino acids can be seen as marginals of a probability distribution

$$f_i(\sigma_a) = \sum_{\sigma_k : k \neq a} p(\sigma_1, \sigma_2, \ldots, \sigma_n) \tag{2}$$

$$f_{ij}(\sigma_a, \sigma_b) = \sum_{\sigma_k : k \neq a,b} p(\sigma_1, \sigma_2, \ldots, \sigma_n) \tag{3}$$

where the probability p is defined as

$$p(\sigma_1, \ldots, \sigma_n) = \frac{1}{Z} \exp\left(-U(\{\sigma_i\})\right). \tag{4}$$

Starting from our alignment with n residues and M sequences of length L, we can compute the empirical frequencies $\tilde{f}_{ij}(\sigma_a, \sigma_b)$ and $\tilde{f}_i(\sigma_a)$. For $M \to +\infty$, the empirical frequencies $\tilde{f}_i, \tilde{f}_{ij}$ will match the theoretical distributions f_i, f_{ij}. For a realistic situation, using the empirical distribution we will have finite-size effects, given by the finite number of sequences available. To overcome this issue, we reweight the empirical frequencies with the appropriate pseudocounts [25]. These pseudocounts are weighted on the distribution of residue types (via the weight x), on the distribution of residue types in the considered alignment (via the weight y), and on the distribution of residue types in a specific pair of positions in the alignment (via the weight z), namely

$$\begin{aligned} f_i(\sigma_a) &= \frac{1}{M_e(x+y+z+1)} \times \left[\tilde{f}_i(\sigma_a) + x\frac{M_e}{q} + y\frac{\sum_j \tilde{f}_j(\sigma_a)}{L} + z\tilde{f}_i(\sigma_a) \right] \\ f_{ij}(\sigma_a, \sigma_b) &= \frac{1}{M_e(x+y+z+1)} \left[\tilde{f}_{ij}(\sigma_a, \sigma_b) + x\frac{M_e}{q^2} + \frac{y}{L^2 M_e} \sum_{kl} \tilde{f}_k(\sigma_a)\tilde{f}_l(\sigma_b) + \frac{z}{M_e}\tilde{f}_i(\sigma_a)\tilde{f}_j(\sigma_b) \right], \end{aligned} \tag{5}$$

where q is the number of residue types (here we have 21 types: the 20 amino acids and the gap), $M_e = \sum_s 1/m_s$ is an effective number of sequences and m_s is the number of sequences with similarity larger than 70%. In this work we fixed $x = 0.5, y = 0.1$, and $z = 1.0$

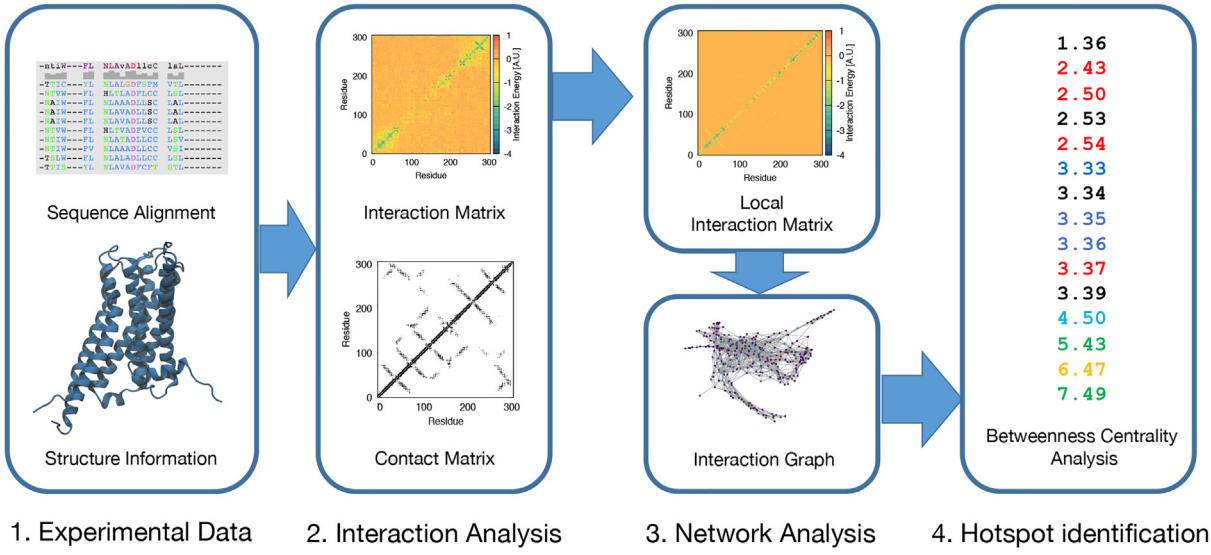

Fig. 1. Schematic representation of the workflow. We employ available sequence alignments and structures to build a coevolution-based interaction matrix that we refine using a contact map, building a network that contains spatial and interaction information about the protein of interest. Hotspots are finally identified by means of the analysis of the betweenness centrality of every node, that are subsequently labeled based on the data available in the literature.

Table 1
Human Class A GPCRs experimental and predicted structures. The homology models were obtained from GPCRdb [23] if the sequence identity between target and template was >50%. Some of the human chemokine receptors structures are only in the inactive state and we also analyzed models for inactive conformations that did not have an experimental structure.

Name	Species	UNIPROT	Active	Inactive
Rhodopsin	Human	OPSD_HUMAN	6CMO	(98%)
Cannabinoid-1	Human	CNR1_HUMAN	6N4B	5U09
Cannabinoid-2	Human	CNR2_HUMAN	(66%)	5ZTY
Muscarinic M1	Human	ACM1_HUMAN	6OIJ	5CXV
Muscarinic M2	Human	ACM2_HUMAN	4MQS	3UON
Muscarinic M4	Human	ACM4_HUMAN	(91%)	5DSG
β_2-Adrenoreceptor	Human	ADRB2_HUMAN	4LDE	2RH1
Adenosine A1	Human	AA1R_HUMAN	6D9H	5UEN
Adenosine A2A	Human	AA2AR_HUMAN	5G53	5NM4
δ-Opioid	Human	OPRD_HUMAN	(82%)	4N6H
μ-Opioid	Human	OPRM_HUMAN	5C1M	4DKL
κ-Opioid	Human	OPRK_HUMAN	6B73	4DJK
NOP Receptor	Human	OPRX_HUMAN	(77%)	5DHH
Serotonin 1B	Human	5HT1B_HUMAN	6G79	(60%)
Serotonin 2A	Human	5HT2A_HUMAN	(83%)	6A94
Serotonin 2B	Human	5HT2B_HUMAN	5TUD	(78%)
Serotonin 2C	Human	5HT2C_HUMAN	6BQG	6BQH
Dopamine 2 Receptor	Human	DRD2_HUMAN	(60%)	6CM4
Dopamine 3 Receptor	Human	DRD3_HUMAN	(57%)	3PBL
Dopamine 4 Receptor	Human	DRD4_HUMAN	(57%)	5WIU
Angiotensin 1	Human	AGTR1_HUMAN	6DO1	4YAY
Apelin Receptor	Human	APJ_HUMAN	(54%)	5VBL
C–C Chemokine 2	Human	CCR2_HUMAN	–	6GPX
C–C Chemokine 5	Human	CCR5_HUMAN	–	5UIX
C–C Chemokine 9	Human	CCR9_HUMAN	–	5LWE
C–C Chemokine 1	Human	CCR1_HUMAN	–	(78%)
C–C Chemokine 3	Human	CCR3_HUMAN	–	(77%)
C–C Chemokine-Like 2	Human	CCRL2_HUMAN	–	(62%)

following Contini and Tiana [13], that tested these parameters for both cytosolic and membrane proteins.

After the reweighting, if we apply a mean-field approximation we can obtain the associated correlation matrix $\mathscr{C}_{ij}$ for the frequencies defined as

$$\mathscr{C}_{ij}(\sigma_a,\sigma_b) = f_{ij}(\sigma_a,\sigma_b) - f_i(\sigma_a)f_j(\sigma_b),$$

and finally obtain the two-bodies interaction energy in Table 2

$$\mathscr{E}_{ij}(\sigma_a,\sigma_b) = (\mathscr{C}^{-1})_{ij}(\sigma_a,\sigma_b) \tag{7}$$

As shown in Refs. [3,12].

This two-bodies term $\mathscr{E}_{ij}(\sigma_a,\sigma_b)$ (that have the form of a 4-dimensional tensor) contains the two-bodies interaction of all the possible residue pairs. Choosing our sequence of interest, we can extract from this tensor the interaction matrix E_{ij} that describes the interaction between all the possible amino acid pairs in the system.

2.3. Network analysis

E_{ij} contains interaction information that is complementary to 3D structural data, because can discriminate the most important energetic contribution of residue that are located in spatial proximity. To integrate spatial and energetic information in a single object, we computed a residue-residue contact map Q_{ij} for every protein structure:

$$Q_{ij} = \begin{cases} 1, & \text{if} |r_i - r_j| \leqslant 10\text{Å} \\ 0, & \text{if} |r_i - r_j| > 10\text{Å} \end{cases} \tag{8}$$

Where $|r_i - r_j|$ is the distance between the C_βs of the i-th and the j-th residues (in the case of glycine, the hydrogen atom in the analogue position).

To insert structural information in E_{ij}, we proceed following Scarabelli et al.[26]. There, the authors perform an Hadamard product (a element-by-element multiplication of the two matrices) between the contact map and the interaction matrix, obtaining the local interaction matrix L_{ij}, namely

$$L_{ij} = Q_{ij} \otimes E_{ij} \tag{9}$$

If the experimental structure considered contains non-resolved residues, one can remove the part of E_{ij} involving such parts.

Now, let us consider each residue of the protein family as a node of a weighted graph $\mathscr{G}$. In this graph, we connect two nodes if the modulus of the interaction obtained from DCA between the respective residues is larger than a threshold value e_{thr} defined iteratively: we start building a network considering the maximum energy of the matrix (in modulus), obtaining an unconnected graph (*i.e.*, a network of isolated nodes except for the two residues with the strongest interaction).[1] At this point, we iteratively lower e_{thr} value until we obtain a connected graph. The maximum value of e_{thr} that still returns a connected graph defines the final interaction network and the connected graph ($\mathscr{G}$) itself.

2.4. Hotspots

The betweenness centrality of a node $B(k)$ in $\mathscr{G}$ reads [27]:

$$B(k) = \sum_{i,j \in \mathscr{G}} \frac{\xi(i,j|k)}{\xi(i,j)} \tag{10}$$

where $\xi(i,j|k)$ is the number of the shortest paths in the graph that connect i and j passing through k, and $\xi(i,j)$ is the number of the shortest paths in the graph that connect i and j.

If the considered node is "central" in the network (*i.e.*, the information flow passes through it to connect different portion of the protein), its betweenness centrality will be larger. We considered

[1] In graph theory a connected graph is a network where, choosing any possible pair of nodes i and j, it is possible to go from i to j via a path defined by the edges of the graph.

residues as hotspots if the betweenness centrality of their associated node is larger than half of the maximum betweenness centrality in all the nodes.

2.5. Application to class A hGPCRs

hGPCRs, with more than 800 members [19], is the largest family of cell-surface receptors. External signals are translated by this family into cell stimuli. A widely used classification system of hGPCRs is the A-F system that is mainly based on their amino acid sequences and functional similarities. This classification identifies six classes, labeled A-F. Class A, also known as the "rhodopsin-like family", is the largest group of hGPCRs [28], which includes hormones, neurotransmitters, and light receptors and accounts for around 80% of hGPCRs. These proteins share a common topological signature, namely seven α-helical transmembrane (TM) domains [29]. The members of the family share also positions of residues directly involved in ligand binding and receptors activation. These include for example positions 3.28, 3.32, 3.33, 3.36, 3.37, 5.39, 6.44, 6.48, 6.55, 7.35 and 7.39 [30–32], functionally conserved along the entire class A [32,20].[2] Such common structural organization contrasts strongly with the agonists' structural diversity, from subatomic particles (a photon), to small molecules, up to peptides and even proteins [29]. Agonist binding to class A hGPCRs triggers receptor activation. The residues involved in the activation extend from the binding cavity, [20,34–37] to the intracellular side of the receptor. Activation lead to binding of a cognate proteins, e.g. G-protein and β-arrestins, and finally downstream signaling pathways. However, these proteins do not simply switch between alternative agonist-bound and inactive forms in this process. They rather adopt a series of intermediate states -likely represented by an ensemble of conformations [38]- influenced not only by agonist binding, but also by other receptors, signaling and regulatory proteins, by post-translational modifications, and by environmental cues [39].

The input of the workflow (Fig. 1) consists of sequence alignments and of experimentally solved (PDB) structures of vertebrate class A GPCRs.[3] We considered the vertebrates GPCRs for building up the evolutionary history of the family because, out of vertebrate species the classification in subfamilies is more difficult and not always accurate [19].

The subclass sequences were downloaded from the Uniprot database [22]. The reviewed sequences were firstly chosen (2514 sequences). New sequences from the unreviewed data set were then manually added. All the resulting curated sequences (5,000) were aligned using the Promals3D web-server [24]. We used the default parameters of the server. The MSA obtained by using this program satisfies all the class A hGPCR features, a set of highly conserved residues in each of the transmembrane helices [30], that gives rise to the Ballesteros-Weinstein nomenclature [33] (see footNote 3). The alignment was aided by 50 experimental or predicted structures belonging to 28 different human class A GPCRs, both in the active and inactive states (see Table 1). In all the cases, the structure was not resolved in its entirety: parts of the sequence (typically the intracellular loop and the C- and N-termini of the chain) was missing in the experimental structure. To match structural/sequence information (and matrices dimension), we removed the parts of the E_{ij} that involved unresolved residues. Next, we built the local interaction network and computed the betweenness centrality of every residue. As mentioned in the description of the method, in this phase we used again the structural information of Table 1. Several hotspot positions underwent site-directed mutagenesis experiments (see Tables 2 and 1 SI and references within). Many mutants have a lower ligand activity or prevent activation (https://gpcrdb.org) [23]) or are linked to disease. Residues identified as hotspots across 30% or more hGPCRs have a documented biological function (Table 2), such as: belonging to the ligand binding site (i), or to the micro-switch network of activation (ii) or being located within the allosteric Na^+ binding cavity (iii). Not all the hotspots listed in Table 2 are present in all the structures in Table 1 (the number of hotspots per single structure ranging from 2 to 26, see Table 2 SI). In particular, for some identified hotspots we can infer their functional role via a direct comparison with other members within the same family. Some examples are briefly discussed below,

2.5.1. Ligand binding

Our protocol is able to capture residues with a fundamental role in selectivity, ligand binding affinity [30] and in dynamical events underlying hGPCR activation (see Table 2; [31,32,20]). For example, it identifies the conserved hotspots involved in ligand binding 3.32, 3.37 and 6.48 [40,30,32,20] (Tables 2 and 1 SI for references). The first residue plays a role for ligand charge detection [32]. It is an aspartic acid in 22% of the class A hGPCRs, interacting with amines or with other positively charged groups [32]. The second residue is involved not only in ligand recognition but also in receptor activation [41]. The last one, is a tryptophan residue in more than 77% of the class A subfamily. This position is well known in literature, since it is a hub involved either in ligand detection and as forming the 'toggle switch' involved in receptor activation (see below) [42].

2.5.2. Micro-switches

The so-called "micro-switches" are small groups of residues that undergo conformational changes during receptor activation and are mechanistically involved in the activation of GPCRs [43,42,44]. Those include: (i) D[E]R3.50Y in helix III which is a common motif that forms the 'ionic lock', during inactivation, (ii) NP7.50xxY in helix VII which plays an important role in proteins' conformational changes upon activation [42]. The link between this region and the binding cavity is the 'toggle switch' formed by positions 6.44, 6.48, 3.40, 5.50: upon ligand binding, position 3.40 rotates and locates between 6.48 and 6.44. This, induce a conformational change of the "hydrophobic barrier", located below the "toggle switch" that includes positions 2.43, 2.46, 3.43, 3.46, 6.36, 6.37 and 6.40 [45]42. The conformational change is important for receptor activation [30,31]. All the residues involved in this complex mechanism were detected as hotspots in one single run of our protocol (Table 2).

2.5.3. Na^+ binding cavity

An allosteric binding cavity for a partially hydrated Na^+ ion is conserved across class A hGPCRs, excluding the opsins [46]. The hydrated Na^+ is bound in the middle of 7TM helices bundle, it stabilizes the inactive state and reduces basal G-protein activity [46]. D2.50 (90% conserved as Asp) directly coordinates the Na^+ ion, N1.50 (97% Asn), S3.39 (75% Ser), N7.45 (70% Asn), S7.46 (66% Ser), N7.49 (75% Asn), and finally Y7.53 (89% Tyr) [46] complete the coordination of the ion. The protocol identifies, as hotspots, D2.50, S3.39 and 7.49 across class A subfamily members, except for the opsins, consistently with experiments [46] (Table 2 SI).

[2] From here on, we will use the Ballesteros-Weinstein (or generic) numbering scheme [33] commonly used for class A GPCRs. Within this framework, the first number indicates the helix and the second number indicates the residue position with respect to the most conserved residue in that helix (x.50). For example, position 3.52 refers to a residue in helix 3, two positions after the most conserved residue, the 3.50.

[3] We consider only non-olfactory class A GPCRs as in [19].

Table 2
Details of the hotspots. For every hotspot identified, we highlight the state (active or inactive) of the hGPCR where the residue was identified, the presence of a documented function, interaction with a ligand, the existence of a mutant or variant in the GPCRdb, and the amino acid consensus. Hy indicates general hydrophobic residues; Ha, Hydrophobic aliphatic; Hb, hydrogen bonding; Sm, small. For references of the experimental data, see [11] to [100] of SI.

Position	Active	Inactive	Function	Ligand contact	Variant GPCRdb	Mutant GPCRdb	Prop Consensus sample	Prop Consensus A-Family	Position	Active	Inactive	Function	Ligand contact	Variant GPCRdb	Mutant GPCRdb	Prop Consensus sample	Prop Consensus A-Family
1.31	0	1				1	\	\	**4.49**	2	0					Hy(81%)	Hy(85%)
1.36	0	1					Hy(81%)	Hy(72%)	**4.50**	2	4			1		*W(100%)*	W(96%)
1.39	0	1				1	Hy(74%)	Hy(86%)	**4.54**	0	1					Hy(81%)	Hy(87%)
1.43	3	0			1		Hy(78%)	Hy(81%)	**4.56**	0	2					Hy(96%)	Hy(87%)
1.46	4	0					\	Sm(79%)	**4.57**	2	0			1		Sm(85%)	Sm(69%)
1.48	0	1					Hy(85%)	Hy(90%)	**4.58**	1	0					\	\
1.50	1	0					*N(100%)*	Hb(100%)	**4.61**	1	1				1	Hy(78%)	Hy(74%)
1.53	1	2			1		*V(96%)*	Ha(90%)	**5.39**	0	2				1	Ha(56%)	Hy(80%)
1.54	2	3					Hy(100%)	Hy(98%)	**5.40**	0	1					Gap(99%)	Gap(99%)
2.38	1	0					Hy(89%)	Hy(73%)	**5.43**	5	6			1	2	Hy(96%)	Hy(58%)
2.39	1	0					Hbo(81%)	Hb(55%)	**5.50**	0	1					Hy(100%)	Hy(95%)
2.42	0	2					Hy(100%)	Hy(97%)	**5.54**	1	2					Hy(93%)	Hy(87%)
2.43	2	11			3		Ha(100%)	Hy(97%)	**5.56**	1	0					Hy(85%)	Hy(80%)
2.46	0	1					*L(100%)*	Ha(98%)	**5.57**	6	6			2		Hy(81%)	Hy(84%)
2.47	2	0			1		*A(100%)*	Ha(81%)	**5.58**	1	1					*Y(100%)*	Hb(89%)
2.48	0	1					Hy(96%)	Hy(94%)	**5.61**	3	0					Ha(85%)	Hy(88%)
2.49	0	1					Sm(96%)	Sm(87%)	**5.62**	1	1			2	1	Hy(89%)	Hy(82%)
2.50	10	15			3	9	*D(96%)*	Hb(98%)	**5.63**	1	0					Hb(63%)	Hb(54%)
2.51	1	1			3		Hy(96%)	Hy(93%)	**5.66**	0	1					Hb(78%)	Hb(60%)
2.54	0	1			1		\	Hy(59%)	**5.68**	0	2					Hb(89%)	Hb(57%)
2.56	2	5			2	6	\	Gap(62%)	**6.26**	1	0					\	Hb(50%)
2.57	5	9				4	Hb(52%)	Hy(56%)	**6.27**	2	0			1	1	\	\
2.59	1	1					Hy(85%)	Hy(92%)	**6.29**	1	0					Hb(74%)	Hb(59%)
2.60	4	2					Hy(96%)	Hy(94%)	**6.30**	1	7			1	4	Hb(89%)	Hb(80%)
2.61	2	1			1	1	Hy(63%)	Hy(52%)	**6.31**	1	1			1		Hb(85%)	Hb(74%)
2.62	0	1					Hy(74%)	Hy(72%)	**6.33**	2	1					Ha(100%)	Hy(71%)
2.63	0	1					Hy(63%)	Hy(75%)	**6.34**	0	1					Hy(59%)	Hy(72%)
2.64	0	1			1	1	Hb(70%)	Hb(65%)	**6.36**	2	3				1	Ha(67%)	Hy(59%)
3.22	0	1					Hb(70%)	Hb(64%)	**6.37**	1	2					Ha(96%)	Hy(91%)
3.25	1	2			1		*C(93%)*	C(88%)	**6.38**	0	3			2		Hy(74%)	Hy(78%)
3.26	1	0				1	Hb(74%)	Hb(83%)	**6.39**	2	0				1	Ha(89%)	Hy(88%)
3.28	2	4			1	8	Hy(85%)	Hy(70%)	**6.40**	4	6			1		Ha(100%)	Hy(91%)
3.29	2	2			1	2	Ha(67%)	Hb(47%)	**6.41**	0	1					Hy(100%)	Hy(91%)
3.30	1	4			1		Sm(78%)	Hy(72%)	**6.44**	2	2			1	1	Hy(100%)	Hy(92%)
3.31	1	1			1		Hy(85%)	Hy(91%)	**6.45**	1	4					Hy(96%)	Hy(92%)
3.32	3	4					Hb(74%)	Hy(56%)	**6.47**	3	4				3	Hy(78%)	Hy(86%)
3.33	4	2			2	3	Hy(74%)	Hy(64%)	**6.48**	2	6				2	*W(96%)*	Hy(91%)
3.34	3	4					Hy(89%)	Hy(82%)	**6.49**	1	1					Hy(67%)	Hy(78%)
3.35	8	16			6	8	\	Hb(55%)	**6.50**	1	1					*P(100%)*	P(99%)
3.36	7	6			6	1	Hy(67%)	Hy(74%)	**6.55**	2	1				2	Hy(56%)	Hy(50%)
3.37	17	16			2	11	Hb(81%)	Hb(62%)	**6.65**	0	1			1		Gap(81%)	Gap(97%)
3.38	1	0			1	1	Sm(85%)	Sm(66%)	**7.28**	0	1					Hb(56%)	\
3.39	16	15			4	4	Hb(85%)	Hb(83%)	**7.34**	0	1					Hy(93%)	Hy(83%)
3.40	5	2			1		Ha(93%)	Hy(89%)	**7.35**	1	3				2	Hy(93%)	Hy(61%)
3.41	10	11			2	2	Hy(89%)	Hy(90%)	**7.37**	1	2					Hy(100%)	Hy(93%)
3.42	6	10			2	2	Hb(67%)	Hy(51%)	**7.39**	1	1			1		Hb(67%)	Hy(57%)
3.43	1	2				1	Ha(96%)	Hy(97%)	**7.45**	0	2			1	1	Hb(93%)	Hb(93%)
3.44	4	3			2	1	Hy(81%)	Hy(60%)	**7.46**	0	1					Sm(96%)	Sm(87%)
3.45	5	6			2		Hy(96%)	Hy(92%)	**7.47**	4	2				1	Hy(67%)	Hy(81%)
3.46	1	3			1		Ha(93%)	Ha(97%)	**7.48**	3	1					Hy(100%)	Hy(97%)
3.47	3	7			1	1	Hb(74%)	Hb(65%)	**7.49**	8	8			1	1	N(100%)	Hb(98%)
3.50	8	16			6	1	*R(100%)*	Hb(98%)	**7.53**	1	0					Y(100%)	Hr(93%)
3.51	0	1					*Y(100%)*	Hr(85%)	**7.55**	1	0			1		Hy(89%)	Hy(94%)
3.54	0	1					Hy(100%)	Hy(97%)	**7.56**	2	1					Hy(78%)	Hy(75%)
4.35	0	1					Gap(100%)	Gap(100%)	**8.47**	1	4			2		Hbo(76%)	Hb(69%)
4.42	0	2				1	Ha(70%)	Hy(70%)	**8.54**	1	0					Hy(100%)	Hy(92%)
4.45	1	0					Ha(59%)	Hy(83%)	**8.52**	0	1					Hb(89%)	Hb(82%)
4.46	0	1					Hy(67%)	Hy(83%)	**8.57**	0	1			1		Hy(93%)	Hy(76%)

6+	8+	>50%	GPCR	Contact in all trajectories/conserved in all GPCR subfamilies
2->4	4->7	>40%	A family	Contact in a majority of trajectories/conserved in two GPCR subfamilies
1	2->3	>30%	Rhodopsin like	Contact in a minority of trajectories/conserved in one GPCR subfamily
0	1	>20%	More than one	No contact
	0	>10%	One	
		0-9%	None	

3. Conclusions

We have presented an approach able to capture most of the coevolution-related events relevant for the function of a very sparse protein family or subfamily (Fig. 1). The protocol provides information not only on residues spanning along the full-length, but also on different activation states of the receptors. In contrast to previous approaches, we make use of structural information.

Application to human class A hGPCRs structures shows that, from a sparse family multiple sequence alignment, we were capable of extracting all the residues known to be involved in the different aspects of the receptor activity that were previously identified [20,21]. These include i) all the position within the binding cavity with a conserved functional role; ii) residues forming the activation microswitches and iii) residues forming the Na^+ allosteric binding cavity and those that were found to be mutated in correlation with disease. Importantly, the method was able to capture the functional role of all the residues in one single shot.

Our approach is totally general and can easily be extended to other subfamilies of GPCRs for which experimental structures are available.[4] As an example, we cite here the pentameric ligand-gated ion channels (pLGICs) protein family, that mediate fast neurotransmission in the nervous system [47,48]49. These are evolutionary correlated, they share a common architecture that consists of three distinct domains, and they exist in at least three distinct functional states [47,48,50]. By exploiting the available structural information [47,48], the approach might be able to identify all hotspots across the family in a single run. As soon as a statistical significant number of structure will be available, a more specific analysis on single subfamilies of class A hGPCRs can be readily be performed.

One of the present limitations of the protocol regards the study of the oligomers, because the integration of the structural data removes the interaction between residues that are far away in the single monomer. In the future, we plan to remove this restriction integrating oligomers data coming from experiments.

CRediT authorship contribution statement

Filippo Baldessari: Investigation. **Riccardo Capelli:** Conceptualization, Investigation. **Paolo Carloni:** Conceptualization. **Alejandro Giorgetti:** Conceptualization.

Acknowledgments

The authors thank Guido Tiana for providing the code to compute interaction matrices from sequence alignments. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 785907 (HBP SGA2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2020.05.003.

References

[1] Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. Science 2017;355(6322):294–8.
[2] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proc Nat Acad Sci 2009;106(1):67–72.
[3] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Nat Acad Sci 2011;108(49):E1293–301.
[4] Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PloS One 2014;9 (3): e92721.
[5] Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J Comput Phys 2014;276:341–56.
[6] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proc Nat Acad Sci 2013;110(39):15674–9.
[7] Jones DT, Buchan DW, Cozzetto D, Pontil M. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2011;28(2):184–90.
[8] Skwark MJ, Abdel-Rehim A, Elofsson A. Pconsc: combination of direct information methods and alignments improves contact prediction. Bioinformatics 2013;29(14):1815–6.
[9] Burger L, Van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Computat Biol 2010;6(1): e1000633.
[10] Morcos F, Onuchic JN. The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes. Curr Opin Struct Biol 2019;56:179–86.
[11] Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. Proc Nat Acad Sci 2015;112(44):13567–72.
[12] Lui S, Tiana G. The network of stabilizing contacts in proteins studied by coevolutionary data. J Chem Phys 2013;139(15):10B618_1.
[13] Contini A, Tiana G. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. J Chem Phys 2015;143 (2):07B608_1.
[14] Ovchinnikov S, Kim DE, Wang RY-R, Liu Y, DiMaio F, Baker D. Improved de novo structure prediction in CASP 11 by incorporating coevolution information into Rosetta. Proteins: Struct Funct Bioinf 2016;84:67–75.
[15] Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Zidek A, Nelson A, Bridgland A, Penedones H, et al. De novo structure prediction with deeplearning based scoring. Annu Rev Biochem 2018;77(363–382):6.
[16] Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife 2014;3: e02030.
[17] Marchetti F, Capelli R, Rizzato F, Laio A, Colombo G. The subtle trade-off between evolutionary and energetic constraints in protein–protein interactions. J Phys Chem Lett 2019;10(7):1489–97.
[18] El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. The pfam protein families database in 2019. Nucl Acids Res 2019;47(D1):D427–32.
[19] Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB. The g-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. Molecul Pharmacol 2003;63 (6):1256–72.
[20] Zhou Q, Yang D, Wu M, Guo Y, Guo W, Zhong L, Cai X, Dai A, Jang W, Shakhnovich EI, et al. Common activation mechanism of class A GPCRs, eLife 8. .
[21] van Westen GJ, Wegner JK, Bender A, IJzerman AP, van Vlijmen HW. Mining protein dynamics from sets of crystal structures using "consensus structures". Protein Sci 2010;19(4):742–52.
[22] Consortium U. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.
[23] Pándy-Szekeres G, Munk C, Tsonkov TM, Mordalski S, Harpsøe K, Hauser AS, Bojarski AJ, Gloriam DE. GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Res 2017;46(D1):D440–6.
[24] Pei J, Kim B-H, Grishin NV. Promals3d: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 2008;36(7):2295–300.
[25] Altschul SF, Gertz EM, Agarwala R, Schäffer AA, Yu Y-K. Psi-blast pseudocounts and the minimum description length principle. Nucleic Acids Res 2008;37 (3):815–24.
[26] Scarabelli G, Morra G, Colombo G. Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. Biophys J 2010;98(9):1966–75.
[27] Brandes U. On variants of shortest-path betweenness centrality and their generic computation. Social Networks 2008;30(2):136–45.
[28] Attwood T, Findlay J. Fingerprinting g-protein-coupled receptors. Protein Eng, Des Selection 1994;7(2):195–203.
[29] Kobilka BK. G protein coupled receptor structure and activation. Biochimica et Biophysica Acta (BBA)-Biomembranes 2007;1768(4):794–807.
[30] Venkatakrishnan A, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. Nature 2013;494 (7436):185.
[31] Veprintsev D, Venkatakrishnan A, Deupi X, Lebon G, Heydenreich FM, Flock T, Miljus T, Balaji S, Bouvier M, Tate CG, et al. Diverse activation pathways in class a gpcrs converge near the g-protein-coupling region. .
[32] Suku E, Giorgetti A. Common evolutionary binding mode of rhodopsin-like GPCRs: Insights from structural bioinformatics. AIMS. Biophysics 2017;4:543–56. AIMS Press.

[4] In its current version, the approach cannot deal with the olfactory receptors as they are a huge disperse subfamily without a crystal structure.

[33] Ballesteros JA, Weinstein H. [19] integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in g protein-coupled receptors. On: Methods in neurosciences, vol. 25, Elsevier, 1995, pp. 366–428. .
[34] Granier S, Kobilka B. A new era of GPCR structural and chemical biology. Nature Chem Biol 2012;8(8):670–3.
[35] Dalton JA, Lans I, Giraldo J. Quantifying conformational changes in GPCRs: glimpse of a common functional mechanism. BMC Bioinf 2015;16(1):124.
[36] Dror RO, Arlow DH, Maragakis P, Mildorf TJ, Pan AC, Xu H, Borhani DW, Shaw DE. Activation mechanism of the β2-adrenergic receptor. Proc Nat Acad Sci 2011;108(46):18684–9.
[37] Kruse AC, Hu J, Pan AC, Arlow DH, Rosenbaum DM, Rosemond E, Green HF, Liu T, Chae PS, Dror RO, et al. Structure and dynamics of the m3 muscarinic acetylcholine receptor. Nature 2012;482(7386):552.
[38] Gumbart J, Khalili-Araghi F, Sotomayor M, Roux B. Constant electric field simulations of the membrane potential illustrated with simple systems. Biochimica et Biophysica Acta (BBA)-Biomembranes 2012;1818(2):294–302.
[39] Geppetti P, Veldhuis NA, Lieu T, Bunnett NW. G protein-coupled receptors: dynamic machines for signaling pain and itch. Neuron 2015;88(4):635–49.
[40] Sandal M, Behrens M, Brockhoff A, Musiani F, Giorgetti A, Carloni P, Meyerhof W. Evidence for a transient additional ligand binding site in the tas2r46 bitter taste receptor. J Chem Theory Comput 2015;11(9):4439–49.
[41] Ponzoni L, Rossetti G, Maggi L, Giorgetti A, Carloni P, Micheletti C. Unifying view of mechanical and functional hotspots across class A GPCRs. PLoS Computat Biol 2017;13(2): e1005381.
[42] Trzaskowski B, Latek D, Yuan S, Ghoshdastider U, Debinski A, Filipek S. Action of molecular switches in GPCRs-theoretical and experimental studies. Curr Medicinal Chem 2012;19(8):1090–109.
[43] Nygaard R, Frimurer TM, Holst B, Rosenkilde MM, Schwartz TW. Ligand binding and micro-switches in 7tm receptor structures. Trends Pharmacol Sci 2009;30(5):249–59.
[44] Schönegge A-M, Gallion J, Picard L-P, Wilkins AD, Le Gouill C, Audet M, Stallaert W, Lohse MJ, Kimmel M, Lichtarge O, et al. Evolutionary action and structural basis of the allosteric switch controlling β 2 AR functional selectivity. Nat Commun 2017;8(1):1–12.
[45] Tehan BG, Bortolato A, Blaney FE, Weir MP, Mason JS. Unifying family a GPCR theories of activation. Pharmacol Therapeutics 2014;143(1):51–60.
[46] Katritch V, Fenalti G, Abola EE, Roth BL, Cherezov V, Stevens RC. Allosteric sodium in class A GPCR signaling. Trends Biochem Sci 2014;39(5):233–44.
[47] Amundarain MJ, Ribeiro RP, Costabel MD, Giorgetti A. Gabaa receptor family: overview on structural characterization. Future Medicinal Chem 2019;11 (3):229–45.
[48] Amundarain MJ, Viso JF, Zamarreño F, Giorgetti A, Costabel M. Orthosteric and benzodiazepine cavities of the $\alpha 1\beta 2\gamma 2$ gabaa receptor: insights from experimentally validated in silico methods. J Biomol Struct Dyn 2019;37 (6):1597–615.
[49] Jaiteh M, Taly A, Hénin J. Evolution of pentameric ligand-gated ion channels: pro-loop receptors, PloS one 11 (3). .
[50] Miller PS, Smart TG. Binding, activation and modulation of cys-loop receptors. Trends Pharmacol Sci 2010;31(4):161–74.