Structure-based network analysis of an evolved G protein-coupled receptor homodimer interface

Sara E. Nichols,^{1,2,3*} Carlos X. Hernández,^{1,4} Yi Wang,^{1,3,5}
and James Andrew McCammon^{1,2,3}

¹Howard Hughes Medical Institute, University of California, San Diego, La Jolla, California

²Department of Pharmacology, University of California, San Diego, La Jolla, California

³Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California

⁴Department of Applied Mathematics, Columbia University, New York, New York

⁵Department of Physics, The Chinese University of Hong Kong, Hong Kong

Received 29 November 2012; Accepted 17 March 2013

DOI: 10.1002/pro.2258

Published online 30 March 2013 proteinscience.org

Abstract: Crystallographic structures and experimental assays of human CXC chemokine receptor type 4 (CXCR4) provide strong evidence for the capacity to homodimerize, potentially as a means of allosteric regulation. Even so, how this homodimer forms and its biological significance has yet to be fully characterized. By applying principles from network analysis, sequence-based approaches such as statistical coupling analysis to determine coevolutionary residues, can be used in conjunction with molecular dynamics simulations to identify residues relevant to dimerization. Here, the predominant coevolution sector lies along the observed dimer interface, suggesting functional relevance. Furthermore, coevolution scoring provides a basis for determining significant nodes, termed hubs, in the network formed by residues found along the interface of the homodimer. These node residues coincide with hotspots indicating potential druggability. Drug design efforts targeting such key residues could potentially result in modulation of binding and therapeutic benefits for disease states, such as lung cancers, lymphomas and latent HIV-1 infection. Furthermore, this method may be applied to any protein–protein interaction.

Keywords: CXCR4; G protein-coupled receptors; network analysis; statistical coupling analysis; molecular dynamics; dimerization

Additional Supporting Information may be found in the online version of this article.

Sara E. Nichols and Carlos X. Hernández contributed equally to this work.

Grant sponsor: NIH, NSF, Howard Hughes Medical Institute, the Center for Theoretical Biological Physics (CTBP), the National Biomedical Computation Resource (NBCR), Extreme Science and Engineering Discovery Environment (XSEDE); Grant number: TG-MCB110156.

*Correspondence to: Sara E. Nichols, Howard Hughes Medical Institute, University of California, San Diego, 9500 Gilman Dr., M/C 0365, La Jolla, CA 92093-0365.
E-mail: senichols@ucsd.edu

Introduction

G protein-coupled receptors (GPCRs) are the largest class of integral membrane protein receptors in the human genome. While these receptors share a conserved heptahelical fold and interact with a limited number of intracellular proteins, the extracellular recognition of a variety of signals produces diverse physiological results.^{1,2} Oligomeric assembly has been biochemically and biophysically identified, and these studies in conjunction with modeling, suggest it to be implicated in receptor maturation, internalization and could be important in the design of new drugs that act on this class of receptors.^{3–7} The last decade has resulted in new GPCR structures that

greatly improve the understanding of how these receptors bind their respective signaling molecules, change upon activation, and associate with G-proteins.⁸ An active area of research involves addressing homodimerization and heterodimerization and its relationship to activation of the receptors and the role this association plays in signaling. This involves characterizing the molecular determinants of these interfaces.

Chemokine receptors regulate cell migration by detecting extracellular chemokines, cytokines that induce chemotaxis, and are implicated in healthy and pathogenic physiology, most notably human diseases such as cancer and AIDS.⁹ In addition to being a key modulator of signal transduction, CXCR4 and the related CC Chemokine receptor 5, CCR5, are also implicated in the HIV-1 viral entry process. Chemokines can be functional as monomers or higher order oligomers to activate cell migration, suggesting the stoichiometry of these ligands and their receptors play a role in functional regulation.¹⁰ The recently determined CXCR4 dimer interface, as observed by crystallography, is in contrast with previously determined Rhodopsin interfaces, which had been accepted to be a model for the receptor superfamily.⁸ Fig. 1 shows the dimer interface, observed in all five crystal structures of CXCR4, including the T4 lysozyme chimera used to stabilize the crystallization. These crystal structures suggest a potential modulation mechanism and protein–protein interface target for chemical biology tool development and structure-based drug discovery.

In this study, we have identified structurally and evolutionarily important residues for the homodimerization of CXCR4. Two methods were combined to validate crystallographically observed orientations and characterize the dynamics of these molecular interactions. First, important residues that evolve in a correlated fashion were identified, and mapped to the interface observed in all five cocrystal structures of CXCR4. These results support the hypothesis that this interface is biologically relevant. The coevolution relationships were then used to weight networks, built from dynamic sampling of the homodimer complex, to identify modules of spatially connected residues that coevolved in a statistically significant way among homodimerizing chemokine receptors. Our analysis revealed a single module consisting of 11 residues (T117^{3.33}, Y121^{3.37}, L125^{3.41}, I162^{4.51}, A164^{4.53}, L165^{4.54}, L166^{4.55}, L167^{4.56}, I169^{4.58}, V206^{5.45}, and L210^{5.49}). These residues make up a central region on the exterior of the TM bundle at the intersection of helices 3, 4, and 5. They are not only structurally vital but also evolutionarily linked to the homodimerization. The residues also match potential hot spots from computational solvent mapping, and can be further be pursued in a structure-based drug design effort.

Figure 1. Observed crystal structure dimer orientation. One of five similar observed crystal structure homodimers (PDB 3ODU). Lines and cartoon representation of secondary structure are colored according to transmembrane helix number (H1–H7), with hypothetical membrane designated by blue horizontal boundary. Intracellular and extracellular loop regions are colored black, and T4 lysozyme fusion (T4L) gray. Intracellular and extracellular viewpoints are shown with schematized orientations (bottom), to show relative positions of helices to each other at both sides of the membrane.

Important results of interest to the chemokine receptor community, as well as methodology that can be extended to other target systems are presented in this article.

Results and Discussion

Identifying residues that coevolved in chemokine homodimers

Statistical coupling analysis (SCA), a sequence-based approach that quantifies pairwise correlations of amino acid evolution in collections of proteins, was performed on sequences from chemokine receptors that dimerize, relative to sequences from the entire rhodopsin-like class of GPCRs. Details of the analysis can be found in the methods section, and the method has been extensively developed and applied to various systems by Ranganathan and co-workers.^{11–15} SCA quantifies how the evolution of position-specific residue types are correlated and further produced groups of residues, called protein sectors, which are maximally independent sets of coevolved sequence positions. SCA has previously

Figure 2. Statistical coupling analysis is complementary to crystal structure orientations. The most dominant protein sector from independent component analysis, as represented by alpha carbon atom spheres mapped onto the crystal structure. This sector was determined by statistical coupling analysis (SCA), performed on the sequences of CXCR4 Chemokine receptors that dimerize with respect to rhodopsin-like GPCRs. The majority of residues in this sector are found on transmembrane helix 4 (yellow) and 5 (green), but also a small number are found on the extracellular loops (black). Dimer and monomer interface are shown in cartoon representation of secondary structure, colored according to transmembrane helix number (H1-H7).

been used to analyze G-proteins and GPCRs.^{11,16,17} Here we have examined what sectors are identified with respect to dimerization. The most dominant sector found is depicted in Fig. 2, mapped onto the structure of CXCR4 (a full list of sectors can be found in the Supporting Information Table S1). This sector contains the majority of its residues predominantly along TM helices 4 and 5. These sequence-based results support the hypothesis that the parallel, symmetric dimer of CXCR4 observed in all five crystal structures represents a biologically relevant homodimer interface.

In addition to the above connection between sequence and structure at the interface, it is interesting to note that the most dominant sector also contained residues on some of the protein loop regions (Fig. 2, loop regions colored in black), particularly the extracellular loop 2 and 3 (ECL2/3). The dominant sector residues on the beta hairpin of ECL2, found along the inner strand, were observed to interact both with the small molecule antagonist and peptide antagonist (PDB ID 3ODU and 3OE0, respectively). This coevolution of the residues at the crystal dimer interface with residues at the area proximal to the chemokine binding site also suggests a functional connection between the two sites, potentially regulatory in nature.

Dynamic interactions at the interface

Protein sectors provide coevolved sequence residues, and these mapped onto three-dimensional structure,

can provide spatial insight to coevolution. While there are many interactions at the dimer interface observed in CXCR4 crystal structures, these interfaces vary, as seen in Fig. 3. These observed changes in monomer orientation could be due to the location of a T4 lysozyme (T4L) construct, fused with the third extracellular loop to facilitate crystallization of this membrane protein. While the T4L construct is functional,^{8,10} the two T4L fusions interact near the intracellular side of the CXCR4 monomers (Fig. 1). These variations made it difficult to make qualitative connections between coevolving residues and crystal structure orientations and lead us to investigate the dynamic protein contact network. Given the location of the fusion, and the varying extent to which the monomers interact, simulations were carried out with the T4L removed and the ECL3 residues 229 and 230 were reconnected. Three independent, 200 ns molecular dynamics simulations of the highest resolution crystal structure CXCR4 homodimer were carried out to capture the dynamic nature of the interface.

In all three simulations the monomers came closer to each other, bridging the intracellular gap. The conformations sampled by molecular dynamics were clustered and the most commonly observed

Figure 3. Dimer orientations observed in crystal structures and molecular dynamics simulations. Molecular surface representations of dimer interactions from the five cocrystal structures (3OE8, 3OD0, 3OD9, 3OE6, and 3ODU) and a representative structure from the molecular dynamics (MD) simulations are colored by chain.

Figure 4. Structure-based network analysis based on coevolutionary relationship. Statistically significant residues, that highlight coevolutionary information transfer, from the complete structure-based network based on molecular dynamics simulations and sequence coevolution. Hub residues, structurally crucial for the network, are denoted with a square. The network results map onto the CXCR4 crystal structure, preserving the same color-scheme and highlighting hub residues with alpha carbon spheres.

simulated conformation is depicted, in comparison with the crystal structures, in Fig. 3. The monomers made contiguous contacts across transmembrane helices 5, 4, the lower half of helix 3 and the extracellular side of helix 6. While all three simulations equilibrated with interactions spanning the hydrophobic surface of the TM regions, the monomers rotated slightly relative to each other (see Supporting Information Fig. S5). As noted from the crystal structures, the dimer association is driven mostly by hydrophobic interactions. Interestingly, contacts highlighted as substantial in the CXCR4 crystal structures,¹⁰ such as side chain-main chain interactions between Trp195^{5,34} and Leu267^{6,63}, stacking of Pro191 and Trp195^{5,34}, as well as bonding interactions between donor and acceptor Asn192 and Glu268 on opposing monomers fluctuated but generally stayed within proximal distance; similar dynamic results were observed elsewhere.¹⁸ The three simulations indicate that TM helix 3, 4, and 5 contain the most contacts at the interface. Given these observations that deviated somewhat from the observed contacts in the crystal structure, network analysis was then performed to combine dynamic contacts and coevolutionary dimer relationships to reveal residues crucial for dimer interactions.

Interactions emphasized by evolutionary correlation

Initial SCA results highlighted TMs 4-3-5 as evolutionarily important for dimerization, consistent with the crystallographic orientations. To broaden the understanding of dynamic contacts in that region and identify structurally and evolutionarily linked clusters of residues for potential targeting, molecular dynamics simulations were combined with SCA results, using network analysis. Network-based

approaches have been successfully used to discretize interactions in biomolecules for studying various phenomena, including recent work using ensemble based techniques such as molecular dynamics simulations.^{19,20} Given the variation in interactions in available crystal structures, networks were constructed based on significant contacts sampled throughout the MD simulations, and subsequent edge lengths were weighted with the coevolutionary SCA scores.

The network comprises 110 residues, forming over 428 interactions, proximal to the observed crystal structure interface. While these residues are structurally interesting and corroborate the crystal structure interactions over the course of the simulations, filtering for significantly recurrent interactions based on coevolution based edge weights revealed a single module of interest over the course of dynamic sampling, depicted in Fig. 4. This module comprises 11 residues (T117^{3,33}, Y121^{3,37}, L125^{3,41}, I162^{4,51}, A164^{4,53}, L165^{4,54}, L166^{4,55}, L167^{4,56}, I169^{4,58}, V206^{5,45}, and L210^{5,49}), forming 15 interactions, and is robust in structure over different simulation lengths (see Supporting Information Fig. S6). These residues make up a central region on the exterior of the TM bundle at the intersection of helices 3, 4, and 5, as depicted in Fig. 4 in red. This module represents a key structural and coevolutionary area of the dynamic interface.

The module was further analyzed based on interconnectedness, to identify hub residues that represent side-chains essential for higher-order structure. The overall protocol used to identify these important hubs, a novel combination of sequence coevolution and structural analysis from simulations that could be extended to studying other protein-protein interfaces, is schematized in Fig. 5. By

Figure 5. Protocol for integrating sequence-based evolution with structure dynamics. Coevolutionary relationships were derived from multiple sequence alignment and statistical coupling analysis. Structural information about the dimer interface was sampled dynamically using molecular simulations of the crystal structure. The coevolutionary relationships were then used to weight networks built from sampling of the homodimer complex, to identify modules of spatially connected residues that coevolved in a statistically significant way. One statistically significant module, in red, was identified and critically connected hubs indicate vital residues for coevolutionary information flow. These residues are structurally vital and linked to evolution of the receptor.

definition, the majority of interactions within the modules rely on the presence of the hub residues: Y121^{3,37}, L125^{3,41}, L167^{4,56}, V206^{5,45}. Known structural and functional annotations of CXCR4 and related GPCRs, from databases such as UniProt and GPCRdb, validate the importance of these hubs.^{21–23} Most strikingly mutation L125^{3,41}W increases thermo-stability of the CXCR4 homodimer, as well as other GPCRs.^{10,24} While the authors were not aware of functional characterization of Y121^{3,37} in CXCR4, in CC chemokine receptor 1 the Y118A^{3,37} mutation showed compromised trafficking to the plasma membrane.²⁵ Similarly, 164^{4,56} mutation has been linked alteration of functional properties, despite not being proximal to the orthosteric binding site, in the beta 2-adrenergic receptor.²⁶ While the dominant sequence-based residues identified by SCA span the observed crystallographic interface, the structure-based evolutionary module here presents a potential area of interaction, and can likely be targeted with small molecules.

Hotspots are correlated with hub sites at the protein-protein interface

Solvent mapping, as carried with the FTMap server, is a way to identify high-affinity interaction sites on a protein surface and suggests potential novel binding pockets.²⁷ All contact residues do not equally contribute to the binding energy of the interaction, and solvent mapping is an alternative method for identifying the key side-chains involved. Probe clusters are ranked according to an energetic score, and top ranking probes generally correspond to the orthosteric site. In mapping the surface of dominant structures observed during the simulation, we aimed to locate important interaction sites that corroborate the dimer interface and that could be exploited for selective modulation.

Not surprisingly, the orthosteric site was predominantly the top ranked interaction site, indicating this area has the most potential for binding small organic probes. There were additional clusters of occupancy, including one at the crystal structure interface, as shown in Fig. 6. This probe site corresponds to the module identified via our structure-based evolutionary network analysis. This also corresponds to the site of a previously identified residue, for which bulky mutants are known to be associated stabilization.²⁴ One can envision designing a small organic molecule to act in the same way to stabilize the receptor.

Including protein structural information into networks is one way to characterize disease-related states for drug design strategies. Targeting protein-protein contact surfaces, while known to be difficult,²⁸ is advantageous due to avoidance of off-target selectivity problems. The power of network analysis is being harnessed to address this very goal.^{19,29,30}

Figure 6. FTMap probes at the dimer interface. A cluster of probe fragments (shown in stick, various colors) bound to 68 monomer structures from the clustered dimer simulations were found at the interface. Here these probes are superimposed with chain A of crystal structure 3ODU for frame of reference. Residues are structurally vital to the contact network and evolutionarily linked to the homodimerization are in red cartoon and line representation, and overlap with the probe-associated hotspot.

Here we have specifically looked at the CXC Chemokine receptor 4. Taken in conjunction, results from statistical coupling analysis, molecular dynamics based contact networks and solvent mapping suggest that the observed crystal structure interface of CXCR4 is functionally relevant and a module of the interface is particularly important and potentially druggable.

Materials and Methods

Model of CXCR4

The 2.5 Å x-ray cocrystal structure of CXCR4 bound to antagonist (6,6-dimethyl-5,6-dihydroimidazo[2,1-b][1,3]thiazol-3-yl)methyl-*N,N'*-dicyclohexylimidothiocarbamate, 1T1t (PDBid 3ODU) was downloaded from the Protein Data Bank.³¹ All ligands, lipid molecules and T4-lysozyme residues, 900–1201, were removed. Freely available Visual Molecular Dynamics (VMD) tools were used to build the full system.³² A 100 × 70 Å² POPC bilayer patch was constructed using the

VMD membrane plugin. Crystal waters were retained, and two 30 Å layers to hydrate the protein-lipid system. 64 chloride and 50 sodium ions were added to the system to generate 75 mM ionic concentration. The freely available PROPKA web server was utilized to assign protonation states of residues.³³ Tryptophan 125 was mutated back to the wild-type residue leucine. Minimization and dynamics simulations were performed with the freely available scalable package NAMD.³⁴ The system was first equilibrated with an unconnected intracellular loop 3 (IL3), where the T4L was removed. This process involved two parts, equilibration of the lipids (5000 steps minimization and 3ns NVT, with everything but the lipid tails constrained) and then slow equilibration of the protein in the membrane environment (5 ns NPT with protein heavy atoms restrained at 10 kcal/mol/Å², and 5ns NPAT with protein alpha carbon atoms restrained at 5 kcal/mol/Å²). After equilibrating the unconnected protein the two ends of IL3 were slowly pulled together using a harmonic restraint, with gradually decreasing restraint distance. The system was then rebuilt with a peptide bond between residues 229 and 230, and reequilibrated in the same manner as described above. Three independent, 200 ns NPAT simulations, initialized with randomized velocity, were then carried out for this system.

Molecular dynamics simulations

The CHARMM27 and CHARMM36 parameters were used to build the protein, lipid system, respectively.³⁵ Particle Mesh Ewald method was utilized with a 1.0 Å grid spacing.³⁶ All simulations were conducted with time step of 2 fs, while nonbonded and PME calculations were executed every 2 and 4 fs, respectively. Constant pressure of 1 atmosphere was maintained with a Langevin piston, and the temperature of 310 K was maintained with Langevin dynamics (damping coefficient of 1/ps).

Sequence data and multiple sequence alignment (MSA)

All sequences were collected from Universal Protein Resource (UniProt) and National Center for Biotechnology Information (NCBI) Protein database, which are publicly available resources.²³ For the statistical coupling analysis, a set of 717 human rhodopsin-like receptors (class A GPCRs), used to calculate background amino acid frequencies, and a foreground set of 139 human CXC chemokine receptors (47 CXCR1, 6 CXCR2, and 86 CXCR4), which are known to homodimerize,³⁷ were compiled using five iterations of the DELTA-BLAST algorithm (with an E-value cut-off of 10⁻¹⁰⁰).³⁸ Sequences were filtered for uniqueness and sequence errors. A list of accession numbers can be found in the Supporting

Information Tables S2 and S3. Foreground sequences were aligned using the multiple sequence comparison by log-expectation (MUSCLE) algorithm, available as a part of the SeaView 4.3.5 suite.^{39,40}

Coevolution scoring using statistical coupling analysis

Coevolution scoring was based on statistical coupling analysis (SCA), a sequence-based approach that quantifies pairwise correlations of amino acid evolution in collections of proteins that share a common phylogenetic origin, using a modified version of the MATLAB SCA toolbox 5.0 distributed by the Ranganathan research group.^{13,14} Unlike the method described by in Halabi et al., background amino acid frequencies were calculated from the set of class A GPCRs, chosen as an appropriate phylogenetic reference for CXCRs, as opposed to the entire nonredundant protein database.³⁷ Additionally, gap-sites in the foreground set were ignored in the analysis.

The following is a brief overview of constructing an SCA matrix of position correlations. For a more detailed explanation of the SCA method please refer to the Supporting Information of Halabi et al.¹⁴

Position-specific conservation

In this context, positional conservation is defined as the Kullback–Leibler divergence of the probability, p , of observing a frequency, f , for particular amino acid, a , at position i , given a background frequency, q . This probability distribution of p is inherently binomial. When the number of foreground sequences, N , is sufficiently large (greater than 100 is sufficient, for this study N is 184), p can be estimated using the Stirling's Approximation, yielding the following divergence equation:

$$D_i^{(a)} = f_i^{(a)} \ln \left(\frac{f_i^{(a)}}{q_i^{(a)}} \right) + (1 - f_i^{(a)}) \ln \left(\frac{1 - f_i^{(a)}}{1 - q_i^{(a)}} \right)$$

Positional correlations

The correlation between the occurrence of the particular amino acid, a , at position i and another amino acid, b , at position j can be determined using the following expression:

$$C_{ij}^{(a,b)} = f_{ij}^{(a,b)} - f_i^{(a)} f_j^{(b)}$$

where $f_{ij}^{(a)}$ is the joint frequency of observing both a and b , at i and j , respectively.

SCA correlation tensor

The SCA correlation tensor has dimensions $20 \times 20 \times 1 \times 1$, where l is the length of the alignment and 20 is the number of biologically relevant amino

acids. Each element in the tensor is calculated as follows:

$$\Omega_{ij}^{(a,b)} = \left| \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} \right| \left| \frac{\partial D_j^{(b)}}{\partial f_j^{(b)}} \right| C_{ij}^{(a,b)}$$

where $C_{ij}^{(a,b)}$ is the positional correlation between two amino acids, and $\left| \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} \right|$ is the change in position-specific conservation across a particular amino acid, $\mathbf{a}$. One can think of this expression as the inner product between two vectors in an abstract space (position-residue), which quantifies, as mutations occur, how changes in conservation are similar.

Reducing dimensionality to determine positional residue couplings

Each 20×20 matrix in the correlation tensor can be spectrally decomposed into a diagonal matrix, where each value along the diagonal is an eigenvalue from the original matrix. Here, each eigenvalue represents the information content of each eigenmode; however, this decomposition has the property that the first eigenvalue is significantly larger than the others ($\lambda_{ij}^l \gg \lambda_{ij}^c$). In this way, the original matrix can be represented as one singular value, reducing the dimensionality of the tensor to rank two:

$$\hat{\Omega}_{ij} = \lambda_{ij}^l$$

Ultimately this reduction condenses divergence of all residue types, to look at the dominant residue contribution to the specific paired positions, yielding the SCA positional matrix.

Independent component analysis to determine protein sectors

Independent component analysis (ICA) was used to identify maximally independent sets of coevolved sequence positions, termed protein sectors by Ranganathan *et al.*¹⁴ They described spectral decomposition as a way to reduce noise and sort out the different contributions to the correlations. In this way, it provides more informative representation of basic functional units than residue position. ICA is then applied to transform eigenmodes into maximally independent components. ICA historically been applied to signal processing, denoising, as more recently various other biological applications such as microarray and transcription data processing.^{41–43} The top 2% of eigenmodes from the SCA correlation matrix, 17 in total for this study, were taken for optimization. Through an iterative optimization process, the eigenvectors of these top eigenmodes are linearly transformed to maximize their statistical independence from one another. In theory, this linear

transformation of eigenvectors yields well-defined independent sectors as groups of sequence positions projected along the independent components (ICs) of position space.

Network analysis

Determining residue interactions. Using atomic coordinate information from the MD trajectory, distances between atoms from each monomer were calculated:

$$d_{ij} = \|\mathbf{a}_i - \mathbf{b}_j\|$$

where $\mathbf{a}_i$ and $\mathbf{b}_j$ are both coordinate vectors of atoms from distinct selections of atoms, $\mathbf{a}$ and $\mathbf{b}$. An interaction within the network is defined by any two residues, $\mathbf{i}$ and $\mathbf{j}$, that come within 5 Å of each other in a given frame. First, $\mathbf{a}$ and $\mathbf{b}$ were taken to be each monomer within the dimer, respectively, to establish inter-monomer interactions. The process was then repeated within each monomer to establish intra-monomer interactions, whereby selection $\mathbf{a}$ comprises atoms that formed interactions in the previous step, specific to that monomer, and $\mathbf{b}$ comprises atoms of the whole monomer. Self-looping interactions were only allowed for the first step, across the dimer.

Calculating edge lengths. SCA scores, $\hat{U}$, were renormalized by dividing each score by the maximum and subtracting by the minimum, and then converted into edge lengths by applying a negative log transform:

$$L_{ij} = -\ln \left(\frac{\hat{\Omega}_{ij}}{\max(\hat{\Omega})} - \min(\hat{\Omega}) \right)$$

Interactions with infinite-valued edge lengths were ignored in this study.

Information transfer and network rewiring.

The concept of information transfer was used to model the coevolutionary signal transition between residues, i.e., how strongly any pair of residues within the network, N , communicate. The amount of information transfer, I , between any two residues, $\mathbf{i}$ and $\mathbf{j}$, could be obtained by considering all possible paths and weighting them by their edge lengths and self-loops. To calculate information transfer within the network, a matrix, B , is constructed as follows:

$$B_{ij} = 1 - 2L_{ij} + \sum_{j \in N_i} L_{ij}$$

$$B_{ij} = 1 - L_{ij}$$

where L_{ii} represents the edge length of a self-loop, formed by residue $\mathbf{i}$; L_{ij} represents the edge lengths between residue $\mathbf{i}$ and residues $\mathbf{j}$, which within the set, N_i , of residues incident to $\mathbf{i}$. The

matrix, B , is then inverted and used to calculate pairwise matrix of information transfer I :

$$C = B^{-1}$$

$$I_{ii} = (C_{ii} - 2C_{ij} + C_{jj})$$

Based on the coevolution score between contacting residues, coevolution between distant residues could be recalculated in terms of statistical significance. By shuffling the edge lengths a large number of times ($n = 1000$), while keeping the network structure constant, and calculating information transfer values, $\hat{I}_{ij}$, for each residue pair could be generated. These distributions of information transfer were found to follow a double-exponential distribution, as determined via a one-sample Kolmogorov-Smirnov Test. Only residue pairs with P -values less than 10^{-4} were considered further, as these can be considered significantly dependent residues within the network.

Determination of recurrent interactions. Significantly recurring interactions across all networks were determined by applying the Poisson distribution:

$$p_i = 1 - e^{-\lambda} \sum_{k=0}^{c_i} \frac{\lambda^k}{k!}$$

where p_i is the P -value of a particular interaction, i , occurring c times, given an expected count of λ , the average count across all interactions. Only interactions with a P -value less than 10^{-6} were taken to be significant and included in the final network.

Determination of hub residues. The final network was then analyzed using the MCODE algorithm to identify structurally critical residues, termed hubs.⁴⁴ The MCODE algorithm was designed to find densely connected core residues within a network based solely on connection data, as opposed to relying on edge lengths.

Solvent mapping to identify potential interaction hot spots

Trajectory RMSD clustering. The last 150 nanoseconds from each simulation were concatenated and clustered using GROMACS's `g_cluster` tool with a 0.1 nm cutoff for root mean square deviations (RMSD) of superimposed atoms.⁴⁵ Trajectory frames were superimposed using the transmembrane bundle alpha carbons. Of the 85 groups produced from clustering, the top 34 represented roughly 90% of the frames from the trajectory and the cluster

centroids for these groups were selected for solvent mapping. This resulted in 68 representative monomer configurations from the dimer simulation.

FTMap. Solvent probe mapping was completed using the freely available FTMap webserver.²⁷ This software searches the protein surface for regions that bind small, drug-like fragments or probes, and identifies low-energy clusters of these probes. The top 68 monomer configurations from the clustered dimers were submitted using the Protein Surface Mode, with all other user options left at default suggestions. Resulting structures were superimposed and compared to module residues. Probe-residue heavy atom contacts were calculated based on contact lengths of 5.0 Å. Probe occupancy per residue was obtained by normalizing relative to average probe occupancy per residue, over the entire set of structures submitted to the FTMap webserver.

Acknowledgments

The authors would like to thank the anonymous reviewer for his/her helpful suggestion to test the probability distribution. The authors are also grateful for fruitful discussions with Dr. Robert Swift and Alisha Caliman.

References

1. Kolb P, Klebe G (2011) The golden age of GPCR structural biology: Any impact on drug design? *Angew Chem Int Edit* 50:11573–11575.
2. Congreve M, Langmead CJ, Mason JS, Marshall FH (2011) Progress in structure based drug design for G protein-coupled receptors. *J Med Chem* 54:4283–4311.
3. Terrillon S, Bouvier M (2004) Roles of G-protein-coupled receptor dimerization. *EMBO Rep* 5:30–34.
4. Maurice P, Kamal M, Jockers R (2011) Asymmetry of GPCR oligomers supports their functional relevance. *Trends Pharmacol Sci* 32:514–520.
5. Simpson LM, Taddese B, Wall ID, Reynolds CA (2010) Bioinformatics and molecular modelling approaches to GPCR oligomerization. *Curr Opin Pharmacol* 10:30–37.
6. Lohse MJ (2010) Dimerization in GPCR mobility and signaling. *Curr Opin Pharmacol* 10:53–58.
7. Cottet M, Faklaris O, Maurel D, Scholler P, Doumazane E, Trinquet E, Pin J, Durroux T (2012) BRET and Time-resolved FRET strategy to study GPCR oligomerization: From cell lines toward native tissues. *Front Endocrinol (Lausanne)* 3:92.
8. Salon JA, Lodowski DT, Palczewski K (2011) The significance of G protein-coupled receptor crystallography for drug discovery. *Pharmacol Rev* 63:901–937.
9. Wijtmans M, Scholten DJ, de Esch IJP, Smit MJ, Leurs R (2012) Therapeutic targeting of chemokine receptors by small molecules. *Drug Discov Today Technol* 9:e229–e236.
10. Wu B, Chien E, Mol C, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi, CF, Hamel, DJ, Kuhn, P, Handel, TM, Cherezov, V, Stevens, RC (2010) Structures of the CXCR4 chemokine GPCR with

- small-molecule and cyclic peptide antagonists. *Science* 330:1066–1071.
11. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10:59–69.
 12. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 116:417–429.
 13. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
 14. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 138:774–786.
 15. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. *Cell* 147:1564–1575.
 16. Sayar K, Ugur O, Liu T, Hilser VJ, Onaran O (2008) Exploring allosteric coupling in the alpha-subunit of Heterotrimeric G proteins using evolutionary and ensemble-based approaches. *BMC Struct Biol* 8:23.
 17. Dima RI, Thirumalai D (2006) Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Sci* 15:258–268.
 18. Rodriguez D, Gutierrez-de-Teran H (2012) Characterization of the homodimerization interface and functional hotspots of the CXCR4 chemokine receptor. *Proteins* 80:1919–1928.
 19. Di Paola L, De Ruvo M, Paci P, Santoni D, Giuliani A (2013) Protein contact networks: An emerging paradigm in chemistry. *Chem Rev* 113:1598–1613.
 20. Atilgan C, Okan OB, Atilgan AR (2012) Network-based models as tools hinting at nonevident protein functionality. *Annu Rev Biophys* 41:205–225.
 21. Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, Oliveira L, de Vlieg J, Vriend G (2011) GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Res* 39:D309–D319.
 22. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* 31:294–297.
 23. Consortium U (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40:D71–D75.
 24. Roth CB, Hanson MA, Stevens RC (2008) Stabilization of the human beta2-adrenergic receptor TM4-TM3-TM5 helix interface by mutagenesis of Glu122(3.41), a critical residue in GPCR structure. *J Mol Biol* 376:1305–1319.
 25. de Mendonca FL, da Fonseca PCA, Phillips RM, Saldanha JW, Williams TJ, Pease, JE (2005) Site-directed mutagenesis of CC chemokine receptor 1 reveals the mechanism of action of UCB 35625, a small molecule chemokine receptor antagonist. *J Biol Chem* 280:4808–4816.
 26. Green SA, Rathz DA, Schuster AJ, Liggett SB (2001) The Ile164 beta(2)-adrenoceptor polymorphism alters salmeterol exosite binding and conventional agonist coupling to G(s). *Eur J Pharmacol* 421:141–147.
 27. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* 25:621–627.
 28. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450:1001–1009.
 29. Nussinov R, Tsai CJ, Csermely P (2011) Allo-network drugs: harnessing allostery in cellular networks. *Trends Pharmacol Sci* 32:686–693.
 30. Csermely P, Korcsmaros T, Kiss HJ, London G, Nussinov R (2013) Structure and dynamics of molecular networks: A novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* PMID: 23384594.
 31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242.
 32. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 33–38.
 33. Bas DC, Rogers DM, Jensen JH (2008) Very fast prediction and rationalization of pK(a) values for protein-ligand complexes. *Proteins* 73:765–783.
 34. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26:1781–1802.
 35. Brooks BR, Brooks III CL, Mackerell AD, Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kucsera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30:1545–1614.
 36. Darden T, York D, Pedersen L (1993) Particle Mesh Ewald—an N.Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* 98:10089–10092.
 37. Wang JH, Norcross M (2008) Dimerization of chemokine receptors in living cells: key to receptor function and novel targets for therapy. *Drug Discov Today* 13:625–632.
 38. Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL (2012) Domain enhanced lookup time accelerated BLAST. *Biol Direct* 7:12.
 39. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
 40. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27:221–224.
 41. Hyvarinen A, Oja E (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13:411–430.
 42. Engreitz JM, Daigle BJ Jr, Marshall JJ, Altman RB (2010) Independent component analysis: mining microarray data for fundamental human gene expression modules. *J Biomed Inform* 43:932–944.
 43. Yang L, Rau MH, Hoiby N, Molin S, Jelsbak L (2011) Bacterial adaptation during chronic infection revealed by independent component analysis of transcriptomic data. *BMC Microbiol* 11:184.
 44. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2.
 45. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435–447.