Mutations Affecting the Oligomerization Interface of G-Protein-Coupled Receptors Revealed by a Novel De Novo Protein Design Framework

Martin S. Taylor,* Ho K. Fung,* Rohit Rajgaria,* Marta Filizola,† Harel Weinstein,‡§ and Christodoulos A. Floudas*

*Department of Chemical Engineering, Princeton University, Princeton, New Jersey; †Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, New York; ‡Department of Physiology and Biophysics, and §HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York

ABSTRACT Specific functional and pharmacological properties have recently been ascribed to G-protein-coupled receptor (GPCR) dimers/oligomers. Because the association of two identical or two distinct GPCR monomers seems to be required to elicit receptor function, it is necessary to understand the exact nature of this interaction. We present here a novel method for de novo protein design and its application to the prediction of mutations that can stabilize or destabilize a GPCR dimer while maintaining the monomer's native fold. To test the efficacy of this new method, the dimer of the single-spanned transmembrane domain of glycophorin A was used as a model system. Experimental data from mutagenesis of the helix-helix interface are compared with computational predictions at that interface, and the model's results are found to be consistent with the experimental findings. A flexible template was developed for the rhodopsin homodimer at atomic resolution and used to predict sets of three and five mutations. The results are found to be consistent across eight case studies, with favored mutations at each position. Mutation sets predicted to be the most disruptive at the dimerization interface are found to be less specific to the flexible template than sets predicted to be less disruptive.

INTRODUCTION

Compelling evidence indicates that G-protein-coupled receptors (GPCRs) form multimeric complexes with distinct pharmacological and functional properties (for recent review articles, see (1–10)). Although most of this evidence comes from in vitro experiments, recent studies using animal models support a specific role for GPCR oligomerization in vivo and in human pathologies. Accordingly, understanding the basis of protein-protein interaction in GPCR oligomerization will significantly enhance our understanding of the molecular mechanisms underlying GPCR cellular function, with the promise of new and improved therapeutics targeting the complex structures and their mechanisms.

Both computational and experimental efforts have been made to identify the interface of GPCR oligomers, but the specific molecular determinants required for stable protein-protein interaction are still unknown. Rhodopsin is the only GPCR whose native organization in rows of dimers has been demonstrated directly using data from atomic force microscopy (11). Based on these data and the crystallographic monomeric structure of rhodopsin (12), a three-dimensional model of rhodopsin oligomers was proposed (13). Specifically, this model consisted of intradimeric interfaces involving transmembrane (TM) helices TM4 and TM5, and the second intracellular loop, whereas helices TM1 and TM2 and the third intracellular loop were involved in the formation of rhodopsin dimer rows. Although it is possible that some GPCRs use different oligomerization interfaces to achieve functional selectivity, a systematic application of an enhanced correlated mutation-analysis-based approach (14,15) to several rhodopsinlike GPCRs that are known to form homo-oligomers identified TM1 and TM4 most often as putative interfaces of dimerization/oligomerization. Cross-linking studies of substituted cysteine residues in TM4 and TM5 of rhodopsin (16), and in TM4 of dopamine D2 receptor, (17), further supported the involvement of these two helices in the dimerization/oligomerization of GPCRs. The recent demonstration that the putative dopamine D2 receptor homodimerization interface involving TM4 is related to function, and that activation of this receptor requires changes at this interface (18), further justifies the need for an improved understanding of the exact nature of the interaction between GPCR monomers. The final goal is to suggest specific mutations that may affect the interaction between GPCR monomers, and thereby either disrupt cross talk between receptor subunits, or promote specific signaling cascades.

To this end, and to reduce the overwhelmingly large number of mutagenesis and cross-linking experiments that would otherwise be required to obtain the desired structural insight, we developed a novel two-stage framework for computational de novo protein design. The goal of this method is to discover mutations that would stabilize or destabilize the interfaces of a GPCR dimer while maintaining the monomer's native fold to the maximum extent possible. To test the efficacy of this new method, we used the dimer of the single-spanned transmembrane domain of Glycophorin A as a model system, and compared the predictions to experimental

Submitted July 18, 2007, and accepted for publication November 1, 2007.

Address reprint requests to Christodoulos A. Floudas, Dept. of Chemical Engineering, Princeton University, Princeton, NJ 08544. Tel.: 609-258-4595; Fax: 609-258-0211; E-mail: floudas@titan.princeton.edu.

Editor: Costas D. Maranas.

0006-3495/08/04/2470/12 $2.00

doi: 10.1529/biophysj.107.117622

data from mutagenesis studies (19–21). The protein design framework was then applied to the TM4,5-TM4,5 dimer of the prototypic GPCR rhodopsin, and used to predict sets of three and five simultaneous lipid-exposed mutations that are likely to disrupt the dimerization interface of rhodopsin with minimal changes in the structural integrity of the monomers.

THEORY

Novel two-stage framework for de novo protein design

The framework for de novo protein design applied to the investigation of the dimerization/oligomerization interface of rhodopsin involves two main stages. The methodology is outlined in Fig. 1. The first stage of the method, termed the ''sequence selection'' stage, begins with a high-resolution flexible template and uses a distance-dependent force field to select a rank-ordered list of amino acid sequences that are predicted to be of low energy in that template. As indicated in Fig. 1, these sequences are the input to the second stage, the ''fold validation'' stage, in which sequences are selected from a list of sequence positions that were rank-ordered by a criterion of relative level of specificity for the flexible template. This fold validation stage is outlined in Fig. 2.

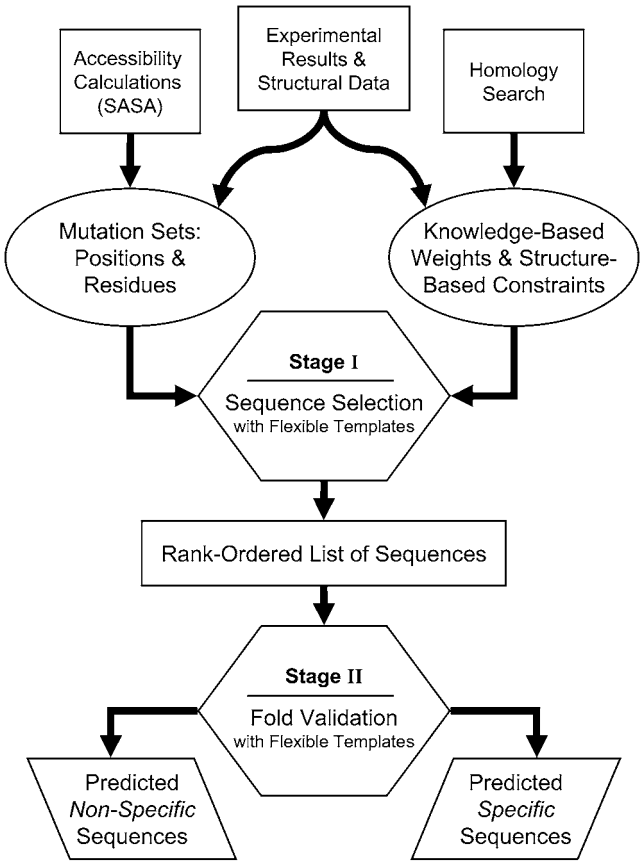

FIGURE 1 Overview of the de novo protein design method. The first stage of the method, the ''sequence selection'' stage, begins with a high-resolution flexible template and uses a distance-dependent force field to select a rank-ordered list of amino acid sequences that are predicted to be of low energy in that template. Fig. 1 shows that these sequences are the input to the second stage, the ''fold validation'' stage, in which sequences are selected from the rank-ordered list of sequence positions that were found to have the highest level of specificity for the flexible template. This fold validation stage is outlined in detail in Fig. 2.

Defining the flexible template

The starting point for any de novo protein design method is the definition of the template or backbone structure. The template is the de novo design method's representation of the desired three-dimensional structure, and therefore appropriate definition is critical. Early methods of de novo protein design assumed a rigid template, with the coordinates of all atoms fixed in space. This assumption was highly convenient because it significantly reduces the complexity of the problem (22). However, it has been observed that this unrealistic constraint renders some problems in de novo protein design untreatable without the use of flexible templates (23–25).

Methods have been created to incorporate template flexibility; two of the most popular are modeling atoms with smaller-than-natural atomic radii and considering a discrete set of templates and sampling among the results (26). Modeling atoms with smaller atomic radii, with typical reductions between 5 and 10%, allows steric overlapping of atoms due to backbone movements. However, this approach has a number of disadvantages including an overestimation of attractive forces between atoms and the possibility of overpacking atoms, particularly in the hydrophobic cores of target molecules (27). Considering a discrete set of templates and sampling among the results alleviates these disadvantages and allows for flexibility to be incorporated through the variations within the sets. Additionally, flexibility can be controlled across different regions of the protein. However, the various templates must somehow be combined into a meaningful three-dimensional structure, and the design problem must be solved for each template considered. This greatly increases the complexity and computational difficulty of the problem. There are several recent reviews of advances in de novo protein design (28–30).

Here, we apply two strategies to incorporate flexibility into the template. Perhaps the most elegant way to incorporate backbone flexibility is to allow variability within the template itself. Following the method of Klepeis and Floudas (26,31–33), we allow for backbone flexibility by incorporating a distance-dependent force field in the sequence selection stage. First, we represent the protein as a matrix of distances between α-carbons. Then, the force field considers pairwise interactions between residues as strictly distance-dependent, allowing rotational and torsional flexibility. Distances between residues are discretized into a set of bins rather than scoring their energy-continuous function. The bin sizes vary between 0.5 and 1 Å, implicitly allowing backbone movements of similar magnitude. It is important to note that, because of the precision required by this method, a high-resolution structure must be used as a starting point.

In addition, when multiple structural models are available, such as the multiple NMR models of the glycophorin A dimer (34), a probability-weighted-average method is used to increase template flexibility by incorporating information from all models (26,27,33).

Developing a flexible template for the rhodopsin dimer

No high-resolution structure of the rhodopsin dimer is yet available. An atomic-level resolution model of a rhodopsin dimer with TM4 and TM5 at the dimerization interface was recently proposed (Protein Data Bank (PDB) identification code 1N3M) that uses atomic force microscopy data (13) and the crystallographic structure of rhodopsin (12). We have recently described the first 45-ns molecular dynamics simulations of this model in an equilibrated unit cell of hydrated palmitoyloleoyl phosphatidylcholine (35). The resulting energy-optimized average structure of the converged interval of these simulations (the last 17.5 ns of the 45-ns simulation) was used to develop a flexible template for the rhodopsin dimer using the strategy outlined in the section above.

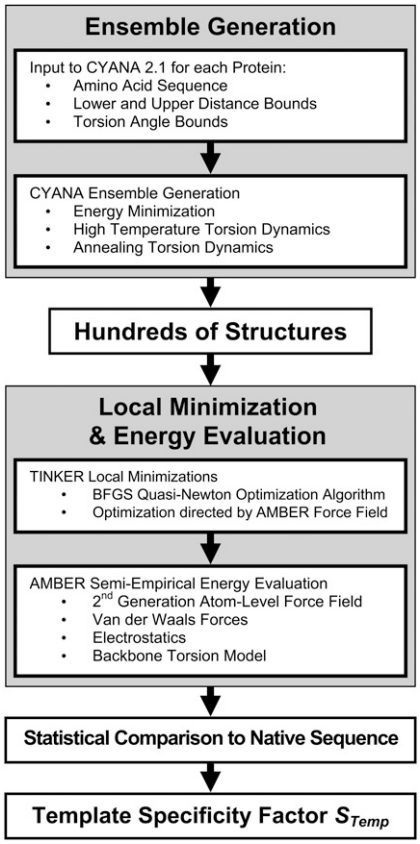

FIGURE 2 Overview of the method for fold validation.

Developing mutation sets

Once a flexible template is defined, an appropriate mutation set is developed. The mutation set dictates which positions are considered for mutation and which residues are allowed in each position. As illustrated in Fig. 1, this selection is influenced by experimental results and the solvent-accessible surface area (SASA). In the general case, all 20 amino acids can be considered at all n positions, and the total combinatorial complexity of the problem is 20^n. However, reducing the search space can be advantageous both to reduce computer time and to increase biological relevance. For example, when redesigning a protein with a particular binding region or catalytically active domain, successful designs often focus on the surrounding regions. Indeed, this was the case for compstatin, a synthetic peptide inhibitor of complement 3, whose activity was increased 45-fold by redesigning residues surrounding the binding loop (31,32,36–39).

Another simple and highly useful method for reducing the search space is to consider only subsets of residues at each position. A popular classification of this kind is to separate the residues based on their environment. This makes sense from a biochemical perspective for a number of reasons. Protein cores are typically composed of tight packings of hydrophobic, nonpolar amino acids. The surface of globular proteins is typically exposed to water, and, accordingly, residues at the surface are generally hydrophilic, polar amino acids. This idea has been used successfully by Harbury and co-workers in their design of α-helical bundle proteins with a right-handed superhelical twist (40), and, recently, by Hecht and co-workers to design a four-helix bundle with a novel fold (41). We introduce this idea in residue selection for our protocol, using the calculated SASA to classify residues into three categories: surface (SASA $>$ 50%), intermediate (20% $<$ SASA $<$ 50%), and core (SASA $<$ 20%). SASA is readily calculated with the program NACCESS (42). In the general case, only hydrophilic residues (GNQHKRDESTP) are considered at surface positions, only hydrophobic residues (AVILMFYW) are considered at core positions, and all 20 amino acids except cysteine are considered at intermediate positions. Cysteine is excluded from the mutation sets because of its ability to cross-link. However, these guidelines can be modified; for example, it may be advantageous to alter the mutation set at a given position if the native residue does not fall into the category predicted by its SASA.

Implementing these approaches, we developed an appropriate mutation set at the dimerization interface of rhodopsin by selecting the following five positions as candidates for mutation: 4.41, 4.48, 4.55, 5.37, and 5.41. We chose these specific positions in TM4 and TM5 because they correspond to the closest symmetric interactions (distance between Cβ atoms $<$11 Å) at the intradimeric interface of rhodopsin, as derived from atomic force microscopy data (13). These positions are identified by the ''generic numbering system'' adopted for GPCRs to refer, comparatively, to structurally cognate receptors (43). Briefly, in this generic numbering scheme, two numbers (N1 and N2) are assigned to amino acid residues in TMs. N1 refers to the TM number, whereas the N2 numbering is relative to the most conserved residue in each TM, which is assigned a value of 50. The other residues in the TM are numbered in relation to this conserved residue, with numbers decreasing toward the N-terminus and increasing toward the C-terminus.

Based on the SASA of native residues in the model (data not shown), two sets of residues were used for the chosen positions of mutations: 1), all residues except cysteine at the helix boundary positions 4.41 and 5.37; and 2), hydrophobic residues plus serine and threonine (AVILMFYWST) at positions 4.48, 4.55, and 5.41. Cysteine was excluded from the mutation sets because previous data show that mutations of lipid-exposed residues to cysteine may cause GPCR cross-linking (16,17), and serine and threonine were added to the standard hydrophobic set because of their prevalence on the lipid-facing surface of membrane proteins.

Energy function

The energy function used in this work is a modified version of the high-resolution (HR) force field developed recently by Rajgaria et al. (44). The HR force field is a knowledge-based C^{α}-C^{α} distance-dependent potential that has been shown to be highly effective in discriminating native protein folds from highly similar sets of decoy structures at high resolution (root mean-square deviation (RMSDs) $<$ 3 Å), and medium resolution (3 Å $>$ RMSD $>$ 10Å) (44). The HR force field contains eight distance bins and considers a contact to be any interaction where the C^{α}-C^{α} distance is $<$9 Å. It is the next generation version of the LKF force field used previously in this method (45). To better model the interactions across the dimerization interface of rhodopsin, a longer-range version of the HR force field was generated and designated the HR-12 force field. The HR-12 force field contains 12 bins, and includes interactions between residues with C^{α}-C^{α} distances up to 13 Å. The HR-12 force field performs similarly to the HR force field on all test cases (data not shown).

The energy measured by this type of HR force field is a distance-dependent approximation of the Gibbs free energy, since it is derived from energy-minimized native conformations in the PDB. Additionally, this type of force field is advantageous for de novo protein design because evaluation of contact energies requires only a table lookup. Hence it is very efficient, allowing for rapid evaluation of many candidate sequences.

The flexible template and mutation set are used to design sets of residues to fit the template structure. Sequence selection is guided by the distance-dependent force field in a mixed integer linear programming model. In the general case, residues are selected to minimize the energy of the structure within the template; the novel sequences generated are designed to be specific to that template. However, to disrupt the dimerization process, we focus on interactions across the dimerization interface and instead maximize the energy, disrupting interactions between the monomer units.

Sequence selection

The sequence selection formulation used in this work is based on the original formulation by Klepeis et al. (31,32). This model was recently improved by

Fung et al. and proven to be totally equivalent to, but computationally more efficient than, the original model (26,27,33). The integer linear programming model then takes the form

$$\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl} \times w_{ik}^{jl}$$

$$\text{subject} \quad \text{to} \quad \sum_{j=1}^{m_i} y_i^j = 1 \quad \forall i$$

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \quad \forall i, k > i, l$$

$$\sum_{l=1}^{m_i} w_{ik}^{jl} = y_i^j \quad \forall i, k > i, j$$

$$y_i^j, y_k^l, w_{ik}^{jl} = 0 - 1 \quad \forall i, j, k > i, l$$

Consider set $i = 1, \ldots, n$ to define the residue positions along the template where n represents the total number of residues. At each position i, the set of mutations is represented by $j\{i\} = 1, \ldots, m_i$, where, for the general case where all mutations are allowed, $m_i = 20 \forall i$. The equivalent sets $k \equiv i$ and $l \equiv j$ are defined so that pairwise interactions can be represented, and $k > i$ is required to ensure that all pairwise interactions are unique. Binary variables y_i^j and y_k^l are introduced to indicate the possible mutations at a given position. That is, when a particular amino acid (j or l) is active at a given position (i or k), variable y_i^j or y_k^l indicate this by taking the value of 1. To ensure that there is exactly one type of amino acid at each position, composition constraints require the sum ofy_i^j be equal to 1 for all positions i. Additionally, there are two sets of RLT constraints that are introduced to reduce the integrality gap. The model minimizes the energy function E_{ik}^{jl} over the entire structure. It is important to note that the energy value then depends on the distance between the α-carbons at the two backbone positions, i and j, as well as the type of amino acids, k and l, at those positions. It should be noted that the binary variables w_{ik}^{jl} can be relaxed into continuous variables as shown by Zhu (46).

To predict sets of mutations that disrupt the dimerization interface of rhodopsin, the objective function is maximized rather than minimized. Accordingly, residues of the highest energy are selected at the dimerization interface.

Increased flexibility and fidelity with multiple structural models

When multiple structural models are available, such as the multiple NMR models of the glycophorin A dimer, a probability-weighted-average method is used to incorporate information from all models (26,27,33). Because of protein flexibility, the C^{α}-C^{α} distances across multiple structural models can differ; for a particular pair of residues, they often span a number of bins. This distribution of distances between residues is reflective of the residues' conformational freedom. We therefore represent the pairwise energy contribution of each pair of residues as a sum of the contributions from all distances spanned, weighted by the probability of finding the residues at each distance.

This energy contribution of the objective function is then written

$$\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d=1}^{b} E_{ik}^{jl} \times \Omega(x_i, x_k, d) \times w_{ik}^{jl},$$

where d is the current distance bin, b represents the total number of distance bins in the force field, and Ω is the weight given to that bin for the current pair of residues, defined as the fraction of structures in which the distance between residues x_i and x_k falls into bin d. This method increases the flexibility of the template by spanning multiple distances for each interaction. At the same time, the fidelity of the model is increased because the probability-weighted distribution is more representative of the interaction between two residues than any single distance can be.

To illustrate this method, imagine a protein for which 20 structural models are available. In these models, the distance between residues α and β falls into bin 2 eight times, bin 3 ten times, and bin 4 two times. The values of Ω for these bins are then 0.4, 0.5, and 0.1, respectively. If the distance-dependent force field E_{ik}^{jl} has the values -6.0, -4.0, and -2.0 units, respectively, for the interaction between α and β over these bins, then the total contribution of the interaction between these residues would be -4.6 units.

Development of structure-based constraints

The mixed-integer linear programming model allows for straightforward incorporation of constraints to increase the biological relevance of the results. As seen in Fig. 1, experimental results and homology information are incorporated in this step, resulting in specifically targeted constraints. For example, charge can be maintained over the entire protein, within specific domains of the protein, or both. The number of simultaneous mutations can be limited, allowing a small number of mutations with maximal effect to be introduced across a large search space. Also, the composition of the protein or a region of the protein can be controlled by imposing minimum and maximum quantities of specific amino acids or groups of amino acids within the design. For example, the number of hydrophobic amino acids in existing β-strands can be bounded to enhance the formation of β-sheets.

Incorporation of knowledge-based weights

We incorporated three distinct sets of knowledge-based weights in the rhodopsin runs to illustrate how such knowledge-based information can be included in the new type of de novo protein design method we present here. Such information can be very useful in improving the specificity of the results. As outlined in Fig. 1, these weights allow knowledge from homology studies to be incorporated into the energy function, supplementing the energy function with structural data focused on the region, or even position, of interest. Incorporation into the energy function is straightforward; the objective function is multiplied by two additional terms $\lambda_i^j \times \lambda_k^l$. The term λ_i^j represents the weight of amino acid j at position i in the protein backbone, and λ_k^l represents the same information for the second residue being considered in the pairwise energy function. Note that for positions not selected for mutation, the weight factors λ are given the value of 1. This prevents bias of some interactions over others based on weight assigned to native residues in non-mutated positions.

One set of knowledge-based input comes from two scales previously described for membrane proteins: the structure surface fraction (SF) scale, and the rhodopsin surface propensity (RhoSP) scale (47). These scales describe what types of residues one might expect to find on the surface of general membrane proteins (SF) and of rhodopsinlike proteins (RhoSP). They are based on the inside/outside-facing distribution of the residues on the template obtained from a bioinformatics procedure described in detail (47). The procedure yields an amino acid property scale (APS) that corresponds to the propensity of residues to be located on the lipid-oriented TM surface in membrane protein. The knowledge-based scale was shown to refine predictions based on conservation criteria alone (47). The APS-based prediction method is available on the web-accessible server ProperTM http://icb.med.cornell.edu/crt/ProperTM/ProperTM.xml.

To increase further the specificity of the structure-based information content, we also developed a position-dependent set of knowledge-based weights for class A GPCRs. These weights represent the probability of finding a residue at a given position across all class A GPCRs. An archive of GPCR sequence information was obtained from the GPCRDB (48) and a database of human class A GPCRs was built from these sequences. In total, there were 545 receptors used. Human receptors were selected to minimize bias for receptors heavily studied across species. Multiple pairwise alignments were then performed using CLUSTAL W 1.83 (49). Taking bovine rhodopsin as a ''gold standard'' for class A GPCRs, each helix of each receptor was aligned with the corresponding helix in bovine rhodopsin. The probability of each residue occurring at each position was then calculated. Note that a residue was tallied only if it was found aligned with one of the residues in bovine rhodopsin; residues aligning with gaps were not tallied.

The resulting weights for the positions mutated in this study are presented in Table 1. Full data for all positions are available in Supplementary Materials, Tables S1–S7).

Method for fold validation

As outlined by the sequence of procedural steps shown in Fig. 1, once a rank-ordered list of sequences has been selected into the template by the sequence selection stage, the sequences are further validated by a second stage, designed to predict the sequence's specificity for the flexible template. In the general case, the same template is used for both stages, and the second stage provides refinement of the results from sequence selection. However, to disrupt the dimerization of rhodopsin, the sequence selection stage maximizes the energy rather than minimizing it. Here, we use the crystal structure of the rhodopsin monomer in the second stage, and the resulting predictions are for sets of mutations that will destabilize the dimerization process of rhodopsin while remaining specific to the native fold of the monomer.

In the two-stage framework developed by Klepeis et al. (31,32), the second stage uses ab initio structure prediction techniques based on the αBB deterministic global optimization solver with an objective function of a full-atomistic force field over the set of independent dihedral angles that describe the configuration of the system (50–64). However, these computations are not currently feasible for proteins the size of rhodopsin. Here, we present an alternative method for fold specificity; this method is outlined in Fig. 2. Briefly, a conformation ensemble is generated by simulated annealing using the CYANA 2.1 package (65,66). The resulting hundreds of structures are subjected to local energy minimizations using TINKER (67) with the AMBER force field (68). The ensembles of structures from candidate sequences are then statistically compared to yield a template specificity factor, S_{Temp}. Details of this procedure are reported below.

First, upper and lower bounds on both the distances between α carbons and the ϕ and ψ angles between residues are extracted from the flexible template. When only one structural model is available, these bounds are defined parametrically, with distances of $\pm 10\%$ and angles of $\pm 25°$. We note that true backbone flexibility is incorporated into this method. The parametrically defined template allows for any possible values of C^{α}-C^{α} distances and dihedral angles.

Then, for both the sequence of interest and the native sequence, an ensemble of random structures (conformers) is generated within the confines of the flexible template. This is done using the CYANA 2.1 software package for NMR structure refinement (65,66). CYANA 2.1 performs annealing calculations that simulate a rapid heating of the protein followed by a slow cooling in which high-temperature torsion dynamics and annealing torsion dynamics are performed. Violations of van der Waals radii and of the flexible template are minimized, thereby minimizing the energy of the target structures. Hundreds of these structures are generated within the confines of the flexible template; for this work we have generated 500 conformers for each sequence. For each conformer in the ensemble, local minimizations are then performed with the TINKER (67) package using the BFGS quasi-Newton optimization algorithm, as guided by gradients in the fully atomistic force field AMBER (68). AMBER is then used to evaluate the potential energy of the final conformer structure. The use of TINKER, AMBER, and CYANA is widespread and documented in the literature, even though there have been reports (69) discussing the inaccuracies of the parameters used within AMBER and suggesting improvements.

To analyze the results, a method similar to that used by Klepeis et al. (31,32) for ensemble comparison in fold validation was employed. First, the mean and standard deviation of both RMSD and AMBER energies were found for the native sequence. Upper bounds on both RMSD and energy were then established; for RMSD, the upper bound was selected as 1.5 standard deviations above the mean, in the energy the upper bound was selected as 2 standard deviations from the mean. A structure is considered to make a contribution to the ensemble only if its energy and RMSD both fall under these upper bounds. This is illustrated in Fig. 3.

The specificity of each mutant sequence to the target template is then calculated relative to the native sequence using a Boltzmann distribution. Both the predicted energy of each conformer and its RMSD from the template structure are used in this calculation.

We define the ''set native'' as the set of all data points from the native sequence that are below both upper bounds, and ''set novel'' as the set of all data points from the novel sequence that meet the same criterion. The template specificity factor, S_{Temp}, is then calculated using Boltzmann probabilities:

$$S_{\text{Temp}} = \frac{\sum_{i\in\text{novel}} \exp[-\beta E_i]}{\sum_{i\in\text{native}} \exp[-\beta E_i]}, \quad \text{where} \quad \beta = \frac{1}{k_B T}.$$

The approximate criterion described here does not claim to be a rigorous calculation of stability or free energy. However, AMBER is designed to quantify the potential energy of a protein in a given three-dimensional conformation, and it provides a good approximation of the enthalpy of the protein when folded into the template. Combined with the large sampling of conformational space around the template, approximating an entropy calculation, the template specificity factor approximates the specificity of the fold within the confines of the flexible template.

In summary, the proposed approach consists of two stages. In the first stage, the sequence selection model minimizes the free energy, which is approximated as a distance-dependent force field derived from existing structures in the PDB. In the second stage, we perform a fold-specificity calculation via two protein-folding calculations, one around the flexible template and the other without template restrictions. Then, based on the ensembles of the generated conformers we rerank the predicted sequences based on the fold specificity. As a result, in stage 1, we aim for better stability, whereas in stage 2 we aim for better specificity, and the proposed approach combines the two in a unique way. This is similar to the ''design-in/design-out'' method described by Koehl and Levitt (70,71).

RESULTS AND DISCUSSION

Glycophorin A: a model system

To test the efficacy of the new method, we compare our results with the experimental data from the Fleming group on the dimerization of glycophorin A (19–21). Their recent studies have probed the relationship between structure, sequence, and stability of the dimerization of glycophorin A, a human erythrocyte protein with a single transmembrane domain

TABLE 1 Class A GPCR position-specific weights

No.	Res	A	V	I	L	F	M	Y	W	C	G	P	T	S	Q	N	H	K	R	E	D
4.41	H	8.2	7.8	0.9	6.7	2.6	2.2	1.7	1.5	3.2	3.3	3.9	5.8	5.6	5.0	4.6	**6.7**	8.0	20.2	0.9	1.3
4.48	F	14.5	20.6	11.9	24.0	**8.1**	5.2	0.8	0.2	4.0	4.4	—	3.6	2.6	—	—	—	—	0.2	—	—
4.55	A	**13.3**	14.5	8.0	30.1	5.1	3.5	1.2	—	1.2	8.6	1.0	5.3	6.6	0.4	1.2	—	—	0.2	—	—
5.37	S	11.1	8.5	10.9	7.2	1.2	1.9	1.9	1.7	0.6	6.0	3.1	5.8	**6.4**	4.7	3.5	1.7	3.1	2.1	16.7	2.1
5.41	Y	8.1	6.6	7.5	11.1	27.5	2.8	**12.6**	1.9	3.2	3.2	0.4	6.4	4.9	0.8	0.6	0.6	0.2	0.8	1.1	0.2

Each value represents the weight of the residue (column) at a specific position (row). Values in bold represent the native residue for bovine rhodopsin at each position. *Res*, residue.

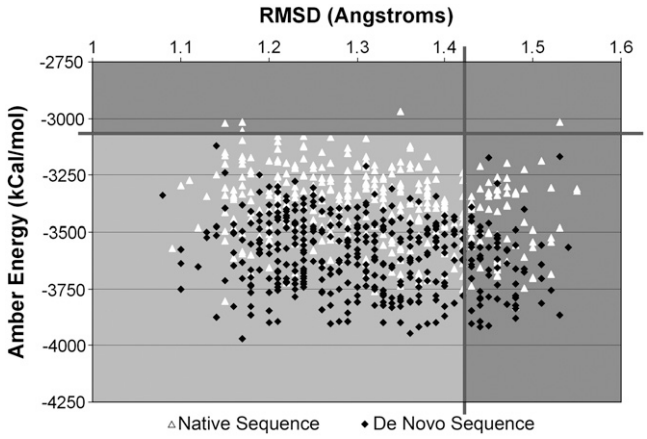

FIGURE 3 Illustration of upper bounds on RMSD and AMBER energy. Thick lines indicate the upper bounds. Data points in shaded regions are not considered in further calculations.

containing a GxxxG motif that is known to form symmetric homodimers. The structure was solved with NMR and the coordinates for 20 models are deposited in the PDB file 1AFO. This information was used to construct the flexible template as described in Methods, and both stages of calculations (see Fig. 1, *Stage I*, to yield rankings and energies, and *Stage II* to yield S_{temp} values) were performed as described in Methods on all reported mutations.

Three case studies had been performed experimentally (19–21), examining three types of mutations. We repeat the studies in silico and compare our results to those found experimentally. We note that the calculations in the two-stage approach presented here relate to the experimentally determined $\Delta\Delta G$ of mutation, but they are not the same. The sequence selection stage minimizes the distance-dependent force field (44) in selection of mutations and the second stage provides a metric for structural specificity. As a result, our approach does not explicitly calculate $\Delta\Delta G$ for the dimer, but our results are consistent with those determined experimentally.

Single alanine mutations

In a first study, single alanine mutations were made along the 75–87 helix fragment of glycophorin A, including both ''lipid-facing'' residues (77, 78, 81, 85, and 86) with side chains pointing away from the dimerization interface, and ''helix-facing'' residues (75, 76, 79, 80, 83, 84, and 87) with side chains pointing into the interface (20). In these experiments, the lipid-facing alanine mutations were found to cause no statistically significant change in the energy of dimerization, with the exception of a moderate stabilizing effect in mutation I85A. In stark contrast with the lipid-facing residues, the helix-facing alanine mutations were found to have moderate to strong destabilizing effects, especially mutation G83A in the GxxxG motif (residues 79–83).

The results from computational studies of these mutants are presented in Tables 2 and 3. In the sequence selection, we find a strong stabilizing effect from mutation I85A. We also note a moderate stabilizing effect of mutation G86A, which does not contradict the experimental results. Whereas the sequence-selection phase fails to identify the stabilizing effect of the GxxxG motif, we find all other mutations destabilizing, in strong accord with the experimental results. Additionally, in the specificity calculations, we also find that all alanine mutations, including those that disrupt the GxxxG motif, produce monomers that are less specific to the template structure than the wild-type sequence.

A number of factors may explain the failure of the sequence-selection phase to identify the GxxxG motif as stabilizing. It has been shown that this complex motif is not required for the dimerization of the glycophorin A transmembrane domain, and that its presence is not sufficient for dimerization (21). Although the function of the glycines in this motif is still not fully understood, proposed mechanisms suggest that they may permit long-range interactions by allowing tighter helical packing. (19,72). Furthermore, where GxxxG motifs occur in the context of the membrane, they may facilitate packing of helical conformations distinct from those seen in soluble proteins (73). Therefore, we propose that the role of these glycines may be different from that played by any glycine used to derive the HR-12 force field. It is also possible that the C^{α}-C^{α} distance-dependent approximation used in the sequence selection phase is not sufficient to model this distinct type of packing.

Single aliphatic mutations along the helix-facing residues

A second study carried out in Fleming's lab (21) explored the requirement of the GxxxG motif (residues 79–83) for

TABLE 2 Single alanine mutations along the dimerization interface of glycophorin A: lipid-facing residues

Stage 1 rank	Stage 1 energy	Stage 2 S_{Temp}	77 ILE	78 PHE	81 MET	85 ILE	86 GLY
3	−733	1.0	ILE	PHE	MET	ILE	GLY
6	−663	0.9	**ALA**	PHE	MET	ILE	GLY
1	−809	0.9	ILE	PHE	MET	**ALA**	GLY
5	−685	0.9	ILE	PHE	**ALA**	ILE	GLY
2	−775	0.8	ILE	PHE	MET	ILE	**ALA**
4	−693	0.8	ILE	**ALA**	MET	ILE	GLY

Each sequence is listed in one row, and the wild-type sequence is identified in the header. Stage 1 energy is in arbitrary units, determined by the distance dependent force field. Higher specificity factor S_{Temp} is better; by definition the native sequence has S_{Temp}=1. Mutated residues are in bold. Wild-type residues at each location are shown in the table header along with the position number.

TABLE 3 Single alanine mutations along the dimerization interface of glycophorin A: helix-facing residues

Stage 1 rank	Stage 1 energy	Stage 2 S_{Temp}	75 LEU	76 ILE	79 GLY	80 VAL	83 GLY	84 VAL	87 THR
3	−733	1.0	LEU	ILE	GLY	VAL	GLY	VAL	THR
6	−679	0.9	LEU	**ALA**	GLY	VAL	GLY	VAL	THR
1	−822	0.8	LEU	ILE	**ALA**	VAL	GLY	VAL	THR
2	−814	0.8	LEU	ILE	GLY	VAL	**ALA**	VAL	THR
8	−667	0.8	LEU	ILE	GLY	**ALA**	GLY	VAL	THR
7	−677	0.7	LEU	ILE	GLY	VAL	GLY	**ALA**	THR
4	−723	0.7	**ALA**	ILE	GLY	VAL	GLY	VAL	THR
5	−695	0.7	LEU	ILE	GLY	VAL	GLY	VAL	**ALA**

Each sequence is listed in one row, and the wild-type sequence is identified in the header. Stage 1 energy is in arbitrary units, determined by the distance-dependent force field. Higher-specificity factor S_{Temp} is better; by definition, the native sequence has S_{Temp}=1. Mutated residues are in bold. Wild-type residues at each location are shown in the table header along with the position number.

glycophorin A dimerization using a series of single mutations of the ''helix-facing'' residues to large aliphatic residues and to glycines. This study also repeats mutations to alanine; however, these are discussed in the previous section and are therefore not repeated. The experiments again find that the GxxxG motif stabilizes the dimer but is not required for dimerization (Fig. 4). As noted previously, our method does not identify the GxxxG motif to be stabilizing in the sequence-selection phase, and therefore mutations that alter these glycines are almost all found to stabilize the dimerization interface. Accordingly, the results segregate those that disrupt the glycines in the GxxxG motif (Table 4) from those that do not (Table 5).

Experimentally, all but two point mutations along the helix-facing residues were found to be destabilizing. Mutation V80L was found to be moderately stabilizing, and mutation I76V was found to have no effect on the energy of dimerization. We find mutation V80L as the best stabilizing mutation in the sequence-selection phase (Rank 1, Energy −763 in Table 5) and in the specificity calculations we find it more specific to the template ($S_{Temp} = 1.1$) than the wild-type sequence ($S_{Temp} = 1.0$). Also, we find mutation I76V to be highly specific to the template. Mutations I76G and I76L are also found to be specific to the template.

Scanning double alanine mutations

In a third experimental study on glycophorin A (19), alanine scan double mutations were performed along the dimerization interface to further probe residue interactions. All double mutants that did not disrupt the glycines in the GxxxG motif were found to be destabilizing in these experiments, and the results were found to be more complex than could be predicted from coupling single mutations. In agreement with these results, we found that all but one mutant were moderately to significantly less specific to the template than wild-type receptor (see Tables 6 and 7, *Stage 2* S_{Temp}). The sole exception to this trend was the double leucine mutant G79LG83L, in which both glycines in the GxxxG motif had been replaced by leucines. Despite the agreement of our second-stage results with experimental data, our method failed to recognize the GxxxG motif as stabilizing. In fact, mutations disrupting both glycines were found to be stabilizing relative to wild-type, as seen by their lower Stage 1 energies.

In the experiments, it was noted that mutations tend not to be additive in their ability to disrupt the dimerization, with a few notable exceptions. In particular, mutation 75A87A was found not to dimerize. Along with a destabilizing result from the sequence selection stage, we find the sequence extremely nonspecific to the template, the least specific among all mutations tested. This is evidenced by the increase in energy relative to the wild-type sequence (Table 6, *Stage1 Energy*) and the extremely low S_{Temp} value of 0.4 (Table 6, *Stage2* S_{Temp}). This is in full agreement with the lack of dimerization found in the experiments.

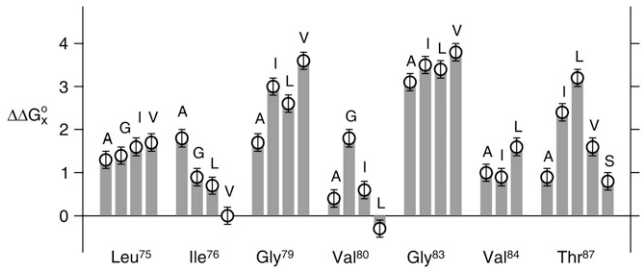

FIGURE 4 Experimentally determined energies of dimerization for single aliphatic mutations. Reprinted from Doura et al. (21).

TABLE 4 Single aliphatic mutations that disrupt the glycines in the GxxxG motif of glycophorin A

Stage 1 rank	Stage 1 energy	Stage 2 S_{Temp}	79 GLY	83 GLY
5	−857	1.3	THR	GLY
7	−843	1.3	GLY	THR
3	−896	1.2	GLY	LEU
1	−933	1.1	LEU	GLY
4	−879	1.1	VAL	GLY
8	−843	1.1	GLY	VAL
11	−733	1.0	GLY	GLY
9	−781	0.9	SER	GLY
10	−778	0.9	GLY	SER
6	−855	0.8	GLY	ILE
2	−898	0.8	ILE	GLY

TABLE 5 Single aliphatic mutations that do not disrupt the glycines in the GxxxG motif of glycophorin A

Stage 1 rank	Stage 1 energy	Stage 2 S_{Temp}	75 LEU	76 ILE	80 VAL	84 VAL	87 THR
4	−747	1.4	LEU	**LEU**	VAL	VAL	THR
12	−719	1.3	LEU	**VAL**	VAL	VAL	THR
1−6	−570	1.1	LEU	**GLY**	VAL	VAL	THR
2	−754	1.1	LEU	ILE	VAL	**LEU**	THR
1	−763	1.1	LEU	ILE	**LEU**	VAL	THR
11	−725	1.0	**VAL**	ILE	VAL	VAL	THR
8	−733	1.0	LEU	ILE	VAL	VAL	THR
7	−736	0.9	LEU	ILE	VAL	VAL	**VAL**
15	−594	0.9	LEU	ILE	**GLY**	VAL	THR
3	−750	0.9	LEU	ILE	VAL	VAL	**LEU**
14	−629	0.8	**GLY**	ILE	VAL	VAL	THR
6	−738	0.8	**ILE**	ILE	VAL	VAL	THR
9	−733	0.8	LEU	ILE	**ILE**	VAL	THR
10	−728	0.8	LEU	ILE	VAL	**ILE**	THR
13	−681	0.7	LEU	ILE	VAL	VAL	**SER**
5	−742	0.7	LEU	ILE	VAL	VAL	**ILE**

Each sequence is listed in one row, and the wild-type sequence is identified in the column heading. Stage 1 energy is in arbitrary units, determined by the distance-dependent force field. Higher-specificity factor S_{Temp} is better; by definition, the native sequence has S_{Temp}=1. Mutated residues are in bold. Wild-type residues at each location are shown in the table header along with the position number.

Dimerization-disrupting mutations of rhodopsin

Generation of the dimer template

Analysis of our results from nanosecond-timescale molecular dynamics simulations (35) of a 1N3M-like intradimeric arrangement of rhodopsin (11,13) revealed a more compact complex structure than had previously been suggested for the rhodopsin dimer (6,13,74). The average structure of the equilibrated portion of these simulations (last 17.5 ns of the 45 ns) was used as a starting point for our novel de novo protein design method.

Selection of positions and mutation sets

As described in the Methods section, the positions investigated for the mutation sets are 4.41, 4.48, 4.55, 5.37, and 5.41. All residues except cysteine are considered for the helix boundary positions 4.41 and 5.37, whereas hydrophobic residues plus serine and threonine (AVILMFYWST) are considered for positions 4.48, 4.55, and 5.41. Additionally, these positions carry a net positive charge due to the histidine at position 4.41. Therefore, we impose a charge constraint, requiring that the mutations maintain the net charge of +1. Because hydrophilic residues are only allowed at two positions, this constraint enforces that exactly one of them carries a positive charge in each solution.

Incorporation of knowledge-based weights

The key criterion of the knowledge-based information incorporated in the application is the probability of finding a particular residue on the outward (lipid)-facing side of the TM helix, rather than on its inward (into the protein bundle)-facing side. This probability has been assessed and quantified from structural data available for membrane proteins (47). The information implicit in the weights from the structure SF method (see (47)) includes the residue distribution between the surface and the interior of membrane proteins, calibrated for the specified regions in the TMs (i.e., the intra- and extracellular parts and the central regions). The RhoSP weights come from an alternative method described in parallel (47), where the crystal structure of bovine rhodopsin and an alignment of 328 rhodopsinlike GPCRs considered to share its structure were used to determine the inside/outside distribution of residues. For each of these criteria, the information obtained from the analysis served to develop an APS that corresponds to the propensity of residues to be located on the TM surface. The weights represent these scales of propensities, as detailed in Table 8.

Use of the class A GPCR position-specific weights provide interesting information regarding the five positions studied here. As expected by their prevalence on the surface of membrane proteins, we see a moderate amount of serine and threonine at all positions. Positively charged residues dominate at position 4.41, providing further support for our enforcement of a positive charge at the interface. Large aliphatic

TABLE 6 Double alanine mutations that do not disrupt the glycines in the GxxxG motif of glycophorin A

	Stage 1 rank	Stage 1 energy	Stage 2 S_{Temp}	75 LEU	76 ILE	79 GLY	80 VAL	83 GLY	84 VAL	87 THR
WT	1	−733	1.0	LEU	ILE	GLY	VAL	GLY	VAL	THR
76A80A	9	−637	0.8	LEU	**ALA**	GLY	**ALA**	GLY	VAL	THR
75A76A	3	−672	0.7	**ALA**	**ALA**	GLY	VAL	GLY	VAL	THR
76A84A	11	−623	0.7	LEU	**ALA**	GLY	VAL	GLY	**ALA**	THR
75A84A	4	−667	0.6	**ALA**	ILE	GLY	VAL	GLY	**ALA**	THR
75A80A	5	−663	0.6	**ALA**	ILE	GLY	**ALA**	GLY	VAL	THR
76A87A	7	−641	0.6	LEU	**ALA**	GLY	VAL	GLY	VAL	**ALA**
80A84A	8	−639	0.6	LEU	ILE	GLY	**ALA**	GLY	**ALA**	THR
80A87A	10	−629	0.5	LEU	ILE	GLY	**ALA**	GLY	VAL	**ALA**
84A87A	6	−660	0.5	LEU	ILE	GLY	VAL	GLY	**ALA**	**ALA**
75A87A	2	−685	0.4	**ALA**	ILE	GLY	VAL	GLY	VAL	**ALA**

Each sequence is listed in one row, and the wild-type sequence is given in the column heading and highlighted in the table. Stage 1 energy is in arbitrary units, determined by the distance-dependent force field. Higher-specificity factor S_{Temp} is better; by definition, the native sequence has $S_{Temp} = 1$. Mutated residues are in bold. Wild-type residues at each location are shown in the column headings, along with the position number.

TABLE 7 Double alanine mutations disrupting the glycines in the GxxxG motif of glycophorin A

	Stage 1 rank	Stage 1 energy	Stage 2 S_{Temp}	75 LEU	76 ILE	79 GLY	80 VAL	83 GLY	84 VAL	87 THR
79L83L	1	−918	1.5	LEU	ILE	**LEU**	VAL	**LEU**	VAL	THR
WT	13	−733	1.0	LEU	ILE	GLY	VAL	GLY	VAL	THR
76A79A	8	−764	0.9	LEU	**ALA**	**ALA**	VAL	GLY	VAL	THR
79A83A	2	−914	0.8	LEU	ILE	**ALA**	VAL	**ALA**	VAL	THR
76A83A	10	−760	0.8	LEU	**ALA**	GLY	VAL	**ALA**	VAL	THR
79A84A	7	−764	0.7	LEU	ILE	**ALA**	VAL	GLY	**ALA**	THR
80A83A	12	−748	0.7	LEU	ILE	GLY	**ALA**	**ALA**	VAL	THR
75A79A	3	−816	0.6	**ALA**	ILE	**ALA**	VAL	GLY	VAL	THR
75A83A	4	−805	0.6	**ALA**	ILE	GLY	VAL	**ALA**	VAL	THR
79A87A	5	−784	0.6	LEU	ILE	**ALA**	VAL	GLY	VAL	**ALA**
83A87A	6	−777	0.6	LEU	ILE	GLY	VAL	**ALA**	VAL	**ALA**
83A84A	9	−761	0.6	LEU	ILE	GLY	VAL	**ALA**	**ALA**	THR
79A80A	11	−756	0.6	LEU	ILE	**ALA**	**ALA**	GLY	VAL	THR

Mutated residues are in bold.

residues dominate at positions 4.48, 4.55, and 5.37. We also note the occurrence of some large hydrophilic residues at 5.37, but we believe that this is due to this position's proximity to the surface of the membrane, which may be slightly different across individual receptors. Aromatic residues are most common at position 5.41, but in keeping the intermediate SASA at this position there is some prevalence of all types of residues.

Analysis of the results

We present eight case studies in total. First, setting an upper limit of three simultaneous mutations, sequences are selected as 1), ''unweighted'' (without the use of any knowledge-based weights); 2), using the class A GPCR position-specific weights; 3), using the SF scale; and 4), using the RhoSP scale. We then repeat these four case studies, allowing up to five simultaneous mutations in each sequence.

TABLE 8 Surface fractions and the rhodopsin SP scale (47)

Residue	Structure SF	Rhodopsin SP scale
A	36	58
C	39	38
D	29	0
E	31	8
F	50	99
G	25	61
H	20	32
I	54	100
K	55	57
L	57	98
M	44	50
N	35	1
P	36	70
Q	33	47
R	39	47
S	27	40
T	37	58
V	53	92
W	60	77
Y	47	84

Note that the scales have been normalized to 100 for easier comparison.

To disrupt the dimerization interface of rhodopsin, the objective function during the sequence selection phase corresponds to the maximization of the energy of the distance-dependent force field. At the same time, we also seek to maintain the structure of the monomer as close as possible to the native fold. Based on initial results from the specificity calculations, we were concerned that total energy maximization would produce sequences that do not fold specifically into the target structure. Therefore, we generate a rank-ordered list of 2,000 sequences for each case and perform the second stage of calculations on solutions 1–40 (the ''most disrupted set'') and 1961–2000 (the ''least disrupted set''). As predicted, the template specificity factor S_{Temp} was much higher in general for the least disrupted set, and only a few solutions from the most disrupted set met the arbitrary minimum value for the template specificity factor S_{Temp} defined in Methods.

The sequence selection calculations were solved using the GAMS modeling language coupled to the CPLEX linear programming solver on an Intel Pentium-4 3.2 GHz workstation. Generation of 2,000 solutions required an average of 6 h. A Beowulf cluster containing 80 nodes of dual 3.0 GHz Intel Xeon processors was used to serially perform the second stage of calculations, and ~36 h of processor time was required for each sequence.

Table 9 presents an illustrative set of results, taken from the case study using the class A GPCR position-specific weights and allowing up to three mutations. We show all results meeting the arbitrary cutoff. We note that in this case study, none of the solutions from the predicted least disrupted set is found to be highly specific to the template structure, with no solution meeting the minimum S_{Temp} value of 5 (maximum value is 1.2: compare to other results below with double- and triple-digit values). This suggests that the GPCR position-specific weights provide key structural information, enhancing the method's ability to find solutions that are structurally destabilizing.

Table 10 presents a summary of the results across all case studies for three mutations. The solutions are all found to be highly specific to the template as measured by the large S_{Temp}

TABLE 9 Illustrative set of results from a case study

S_{Temp}	4.41 HIS	4.48 PHE	4.55 ALA	5.37 SER	5.41 TYR
127	**ILE**	**TYR**	ALA	**ARG**	TYR
120	**TYR**	**TRP**	ALA	**ARG**	TYR
109	**PHE**	PHE	ALA	**ARG**	**ALA**
70	**ARG**	PHE	ALA	**GLY**	**THR**
49	**TYR**	PHE	ALA	**ARG**	**TRP**
25	**ARG**	**TYR**	ALA	SER	TYR
14	**ARG**	PHE	**SER**	SER	**ILE**
13	**ALA**	PHE	**SER**	**LYS**	TYR
8	**GLY**	PHE	ALA	**LYS**	**TRP**
5	**TYR**	PHE	ALA	**LYS**	**MET**

Column entries are the complete results meeting the arbitrary cutoff on S_{Temp} from the case study using the class A GPCR position-specific weights, where up to three mutations are allowed. Note that all sequences are from solutions 1961–2000. Each row represents one sequence. Mutated residues are in bold. The native residues are shown in the column headings.

values. The complete results from all the studies are available in Supplementary Material.

We find a number of general trends across all case studies for three mutations, as summarized in Tables 10 and 11. The histidine at position 4.41 is always mutated. Additionally, there is a trend to shift the positive charge away from this position to position 5.37, natively occupied by a serine. In the sequences with the highest specificity, this position tends to be mutated to arginine, although in some solutions with good specificity it is mutated to lysine as well. Also, position 4.55 is the least mutated; when the wild-type alanine is mutated, it is usually to a serine.

TABLE 10 A compilation of dimerization-disrupting solutions across the three-mutation case studies of rhodopsin

Source	S_{Temp}	4.41 HIS	4.48 PHE	4.55 ALA	5.37 SER	5.41 TYR
UNW	272	**GLN**	PHE	**SER**	**ARG**	TYR
UNW	57	**ALA**	**TRP**	ALA	**ARG**	TYR
UNW*	29	**ARG**	**SER**	**SER**	SER	TYR
UNW	381	**THR**	PHE	**SER**	**ARG**	TYR
UNW	35	**SER**	PHE	ALA	**LYS**	**SER**
GPCRA	120	**TYR**	**TRP**	ALA	**ARG**	TYR
GPCRA	109	**PHE**	PHE	ALA	**ARG**	ALA
GPCRA	127	**ILE**	**TYR**	ALA	**ARG**	TYR
GPCRA	70	**ARG**	PHE	ALA	**GLY**	**THR**
GPCRA	49	**TYR**	PHE	ALA	**ARG**	**TRP**
SF	118	**ALA**	**THR**	ALA	**ARG**	TYR
SF*	19	**ARG**	**TRP**	ALA	SER	**ALA**
SF*	18	**ARG**	**ALA**	ALA	SER	**SER**
SF	8	**PHE**	PHE	ALA	**LYS**	**THR**
SF	59	**ARG**	**TYR**	ALA	**GLN**	TYR
RHO_SP	206	**ASN**	PHE	**SER**	**ARG**	TYR
RHO_SP	190	**GLY**	PHE	ALA	**ARG**	**THR**
RHO_SP*	23	**ARG**	**SER**	ALA	SER	**SER**
RHO_SP	28	**ARG**	PHE	**SER**	SER	**VAL**
RHO_SP	22	**ARG**	PHE	**SER**	SER	TYR

A group of solutions were selected for their high specificity to the template and for their heterogeneity. *UNW*, unweighted; *GPCRA*, class A GPCR position-specific weights; SF, surface-fraction scale; *RHO_SP*, rhodopsin SP scale. Mutated residues are in bold.
*Denotes a solution from the ''most disrupted set''.

TABLE 11 A compilation of dimerization-disrupting solutions across the five-mutation case studies of rhodopsin

Source	S_{Temp}	4.41 HIS	4.48 PHE	4.55 ALA	5.37 SER	5.41 TYR
UNW	403	**GLN**	**TRP**	**SER**	**ARG**	**SER**
UNW*	233	**GLY**	**SER**	**SER**	**ARG**	**SER**
UNW	151	**ALA**	**ALA**	**SER**	**ARG**	**VAL**
UNW*	133	**GLY**	**SER**	**SER**	**ARG**	**ALA**
UNW*	72	**ARG**	**SER**	**SER**	**GLY**	**THR**
UNW*	246	**ARG**	**SER**	**SER**	**ASN**	**THR**
UNW	132	**TYR**	**SER**	**SER**	**ARG**	**VAL**
UNW	121	**ALA**	**SER**	**THR**	**ARG**	**ALA**
GPCRA	224	**GLN**	**TYR**	**TYR**	**ARG**	**TRP**
GPCRA	192	**PHE**	**SER**	**SER**	**ARG**	**THR**
GPCRA*	155	**GLY**	**TRP**	**SER**	**ARG**	**TRP**
GPCRA*	145	**GLY**	**TYR**	**SER**	**ARG**	**TRP**
GPCRA*	107	**GLY**	**SER**	**SER**	**ARG**	**TRP**
SF*	151	**GLN**	**TRP**	**TRP**	**ARG**	**TRP**
SF*	122	**ASN**	**TRP**	**TRP**	**ARG**	**TRP**
SF	95	**ALA**	**SER**	ALA	**ARG**	**MET**
SF	92	**PRO**	**THR**	**TRP**	**ARG**	**SER**
SF	78	**ALA**	**ALA**	**SER**	**ARG**	**TRP**
RHOSP	617	**ASN**	**ALA**	ALA	**ARG**	**ALA**
RHOSP*	382	**ASN**	**SER**	**SER**	**ARG**	**SER**
RHOSP	343	**ASN**	**SER**	**MET**	**ARG**	**THR**
RHOSP	306	**ASN**	**MET**	**THR**	**ARG**	**SER**
RHOSP*	233	**GLY**	**SER**	**SER**	**ARG**	**SER**

A heterogeneous group of solutions were selected for their high specificity to the template. *UNW*, unweighted; *GPCRA*, class A GPCR position-specific weights; *SF*, surface-fraction scale; *RHO_SP*, rhodopsin SP scale. Mutated residues are in bold.
*Denotes a solution from the ''most disrupted set''.

Not surprisingly, when up to five mutations are allowed, we find much more diversity and complexity of solutions, in particular at positions 4.41 and 5.41. However, clear trends are evident. First, there is a near-complete shift of the positive charge from histidine 4.41 to an arginine at position 5.37. This is consistent with the prevalence of this shift in the case studies (above) allowing up to three mutations. Second, we note that serine now dominates at position 4.55, replacing the wild-type alanine residue that was most common in the three-mutation cases. Finally, phenylalanine 4.48 is mutated in each case to tryptophan, tyrosine, or serine.

It is interesting to note that alanine is selected in a number of cases as an amino acid to disrupt the dimerization. Yet in almost all cases, it occurs in a solution from the solution range 1961–2000, indicating that although an alanine mutation may not promote the largest local disruption, it may provide a compromise between disruptions of the dimer and fold specificity. This trend occurs across case studies with both three and five mutations.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

C.A.F. gratefully acknowledges financial support from the National Science Foundation, the National Institutes of Health (grants R01 GM52032 and

R24 GM069736), and the U.S. Environmental Protection Agency. (R832721010). M.F. gratefully acknowledges financial support from the National Institutes of Health (grants R01 DA020032 and R21/R33 DA017976). H.W. gratefully acknowledges financial support from the National Institutes of Health (grants K05-DA00060 and P01-DA012408).

REFERENCES

1. Terrillon, S., and M. Bouvier. 2004. Roles of G-protein-coupled receptor dimerization. *EMBO Rep.* 5:30–34.
2. Javitch, J. A. 2004. The ants go marching two by two: oligomeric structure of G-protein-coupled receptors. *Mol. Pharmacol.* 66:1077–1082.
3. Bai, M. 2004. Dimerization of G-protein-coupled receptors: roles in signal transduction. *Cell. Signal.* 16:175–186.
4. Park, P. S., S. Filipek, J. W. Wells, and K. Palczewski. 2004. Oligomerization of G protein-coupled receptors: past, present, and future. *Biochemistry.* 43:15643–15656.
5. Filizola, M., and H. Weinstein. 2005. The structure and dynamics of GPCR oligomers: a new focus in models of cell-signaling mechanisms and drug design. *Curr. Opin. Drug Discov. Devel.* 8:577–584.
6. Fotiadis, D., B. Jastrzebska, A. Philippsen, D. J. Muller, K. Palczewski, and A. Engel. 2006. Structure of the rhodopsin dimer: a working model for G-protein-coupled receptors. *Curr. Opin. Struct. Biol.* 16:252–259.
7. Bulenger, S., S. Marullo, and M. Bouvier. 2005. Emerging role of homo- and heterodimerization in G-protein-coupled receptor biosynthesis and maturation. *Trends Pharmacol. Sci.* 26:131–137.
8. Pfleger, K. D., and K. A. Eidne. 2005. Monitoring the formation of dynamic G-protein-coupled receptor-protein complexes in living cells. *Biochem. J.* 385:625–637.
9. Prinster, S. C., C. Hague, and R. A. Hall. 2005. Heterodimerization of g protein-coupled receptors: specificity and functional significance. *Pharmacol. Rev.* 57:289–298.
10. Milligan, G. 2006. G-protein-coupled receptor heterodimers: pharmacology, function and relevance to drug discovery. *Drug Discov. Today.* 11:541–549.
11. Fotiadis, D., Y. Liang, S. Filipek, D. A. Saperstein, A. Engel, and K. Palczewski. 2003. Atomic-force microscopy: rhodopsin dimers in native disc membranes. *Nature.* 421:127–128.
12. Teller, D. C., T. Okada, C. A. Behnke, K. Palczewski, and R. E. Stenkamp. 2001. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of g-protein-coupled receptors (GPCRs). *Biochemistry.* 40:7761–7772.
13. Liang, Y., D. Fotiadis, S. Filipek, D. A. Saperstein, K. Palczewski, and A. Engel. 2003. Organization of the G protein-coupled receptors rhodopsin and opsin in native membranes. *J. Biol. Chem.* 278:21655–21662.
14. Filizola, M., and H. Weinstein. 2005. The study of G-protein coupled receptor oligomerization with computational modeling and bioinformatics. *FEBS J.* 272:2926–2938.
15. Filizola, M., W. Guo, J. A. Javitch, and H. Weinstein. 2005. Oligomerization domains of G-protein coupled receptors: insights into the structural basis of GPCR association. *In* Contemporary Clinical Neuroscience: The G Protein-Coupled Receptor Handbook. L. A. Devi, editor. Humana Press, Totowa, NJ. 243–265.
16. Kota, P., P. J. Reeves, U. L. Rajbhandary, and H. G. Khorana. 2006. Opsin is present as dimers in COS1 cells: identification of amino acids at the dimeric interface. *Proc. Natl. Acad. Sci. USA.* 103:3054–3059.
17. Guo, W., L. Shi, and J. A. Javitch. 2003. The fourth transmembrane segment forms the interface of the dopamine D2 receptor homodimer. *J. Biol. Chem.* 278:4385–4388.
18. Guo, W., L. Shi, M. Filizola, H. Weinstein, and J. A. Javitch. 2005. Crosstalk in G protein-coupled receptors: changes at the transmembrane homodimer interface determine activation. *Proc. Natl. Acad. Sci. USA.* 102:17495–17500.
19. Doura, A. K., and K. G. Fleming. 2004. Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. *J. Mol. Biol.* 343:1487–1497.
20. Fleming, K. G., and D. M. Engelman. 2001. Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants. *Proc. Natl. Acad. Sci. USA.* 98:14340–14344.
21. Doura, A. K., F. J. Kobus, L. Dubrovsky, E. Hibbard, and K. G. Fleming. 2004. Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer. *J. Mol. Biol.* 341:991–998.
22. Ponder, J. W., and F. M. Richards. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775–791.
23. Mooers, B. H. M., D. Datta, W. A. Baase, E. S. Zollars, S. L. Mayo, and B. W. Matthews. 2003. Repacking the core of T4 lysozyme by automated design. *J. Mol. Biol.* 332:741–756.
24. Farinas, E., and L. Regan. 1998. The de novo design of a rubredoxin-like Fe site. *Protein Sci.* 7:1939–1946.
25. Ross, S. A., C. A. Sarisky, A. Su, and S. L. Mayo. 2001. Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. *Protein Sci.* 10:450–454.
26. Fung, H. K., M. S. Taylor, and C. A. Floudas. 2007. Novel formulations for the sequence selection problem in de novo protein design with flexible templates. *Optim. Methods Softw.* 22:51–71.
27. Floudas, C. A. 2005. Research challenges, opportunities and synergism in systems engineering and computational biology. *AIChE J.* 51:1872–1884.
28. Butterfoss, G. L., and B. Kuhlman. 2006. Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct.* 35:49–65.
29. Kuhlman, B., and D. Baker. 2004. Exploring folding free energy landscapes using computational protein design. *Curr. Opin. Struct. Biol.* 14: 89–95.
30. Fung, H. K., W. J. Welsh, and C. A. Floudas. 2008. Computational de novo peptide and protein design: rigid templates versus flexible templates. *Ind. Eng. Chem. Res.* 47:993–1001.
31. Klepeis, J. L., C. A. Floudas, D. Morikis, C. G. Tsokos, E. Argyropoulos, L. Spruce, and J. D. Lambris. 2003. Integrated computational and experimental approach for lead optimization and design of compstatin variants with improved activity. *J. Am. Chem. Soc.* 125: 8422–8423.
32. Klepeis, J. L., C. A. Floudas, D. Morikis, C. G. Tsokos, and J. D. Lambris. 2004. Design of peptide analogues with improved activity using a novel de novo protein design approach. *Ind. Eng. Chem. Res.* 43:3817–3826.
33. Fung, H. K., S. Rao, C. A. Floudas, O. Prokopyev, P. M. Pardalos, and F. Rendl. 2005. Computational comparison studies of quadratic assignment like formulations for the in silico sequence selection problem in de novo protein design. *J. Comb. Optim.* 10:41–60.
34. MacKenzie, K. R., J. H. Prestegard, and D. M. Engelman. 1997. A transmembrane helix dimer: structure and implications. *Science.* 276: 131–133.
35. Filizola, M., S. X. Wang, and H. Weinstein. 2006. Dynamic models of G-protein coupled receptors dimers: indications of asymmetry in the rhodopsin dimer from molecular dynamics simulations in a POPC bilayer. *J. Comput. Aided Mol. Des.* 20:405–416.
36. Mallik, B., M. Katragadda, L. A. Spruce, C. Carafides, C. G. Tsokos, D. Morikis, and J. D. Lambris. 2005. Design and NMR characterization of active analogues of compstatin containing non-natural amino acids. *J. Med. Chem.* 48:274–286.
37. Morikis, D., C. A. Floudas, and J. D. Lambris. 2005. Structure-based integrative computational and experimental approach for the optimization of drug design. *Lect. Notes Comput. Sci.* 3515:680–688.
38. Morikis, D., A. M. Soulika, B. Mallik, J. L. Klepeis, C. A. Floudas, and J. D. Lambris. 2004. Improvement of the anti-C3 activity of compstatin using rational and combinatorial approaches. *Biochem. Soc. Trans.* 32:28–32.

39. Sahu, A., B. K. Kay, and J. D. Lambris. 1996. Inhibition of human complement by a C3-binding peptide isolated from a phage-displayed random peptide library. *J. Immunol.* 157:884–891.
40. Harbury, P. B., J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. 1998. High-resolution protein design with backbone freedom. *Science.* 282: 1462–1467.
41. Hecht, M. H., A. Das, A. Go, L. H. Bradley, and Y. Wei. 2004. De novo proteins from designed combinatorial libraries. *Protein Sci.* 13: 1711–1723.
42. Hubbard, S. J., and J. M. Thornton. 1993. ''NACCESS'' Computer Program. Department of Biochemistry and Molecular Biology, University College London, London, UK.
43. Ballesteros, J. A., and H. Weinstein. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* 25:366–428.
44. Rajgaria, R., S. McAllister, and C. A. Floudas. 2006. A novel high resolution Cα-Cα distance dependent force field based on a high quality decoy set. *Proteins.* 65:726–741.
45. Loose, C., J. L. Klepeis, and C. A. Floudas. 2004. A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins.* 54:303–314.
46. Zhu, Y. 2007. Mixed-integer linear programming algorithm for a computational protein design problem. *Ind. Eng. Chem. Res.* 46:839–845.
47. Beuming, T., and H. Weinstein. 2004. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics.* 20:1822–1835.
48. Horn, F., J. Weare, M. Beukers, S. Hörsch, A. Bairoch, W. Chen, Ø. Edvardsen, F. Campagne, and G. Vriend. 1998. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* 26:277–281.
49. Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
50. Klepeis, J. L., and C. A. Floudas. 1999. Free energy calculations for peptides via deterministic global optimization. *J. Chem. Phys.* 110: 7491–7512.
51. Klepeis, J. L., H. D. Schafroth, K. M. Westerberg, and C. A. Floudas. 2002. Deterministic global optimization and ab initio approaches for the structure prediction of polypeptides, dynamics of protein folding, and protein-protein interactions. *In* Computational Methods for Protein Folding. R. A. Friesner, editor. Wiley-Interscience, Hoboken, NJ. 265–457.
52. Adjiman, C. S., I. P. Androulakis, and C. A. Floudas. 1998. A global optimization method, αBB, for general twice-differentiable constrained NLPs. II. Implementation and computational results. *Comput. Chem. Eng.* 22:1159–1179.
53. Adjiman, C. S., I. P. Androulakis, and C. A. Floudas. 2000. Global optimization of mixed-integer nonlinear problems. *AIChE J.* 46:1769–1797.
54. Androulakis, I. P., C. D. Maranas, and C. A. Floudas. 1995. αBB: a global optimization method for general constrained nonconvex problems. *J. Glob. Optim.* 7:337–363.
55. Adjiman, C. S., I. P. Androulakis, C. D. Maranas, and C. A. Floudas. 1996. A global optimization method, αBB, for process design. *Comput. Chem. Eng.* 20:S419–S424.
56. Androulakis, I. P., C. D. Maranas, and C. A. Floudas. 1997. Prediction of oligopeptide conformations via deterministic global optimization. *J. Glob. Optim.* 11:1–34.
57. Adjiman, C. S., S. Dallwig, C. A. Floudas, and A. Neumaier. 1998. A global optimization method, αBB, for general twice-differentiable constrained NLPs. I. Theoretical advances. *Comput. Chem. Eng.* 22: 1137–1158.
58. Klepeis, J. L., C. A. Floudas, D. Morikis, and J. D. Lambris. 1999. Predicting peptide structures using NMR data and deterministic global optimization. *J. Comput. Chem.* 20:1354–1370.
59. Klepeis, J. L., M. Pieja, and C. A. Floudas. 2003. A new class of hybrid global optimization algorithms for peptide structure prediction: integrated hybrids. *Comput. Phys. Commun.* 151:121–140.
60. Klepeis, J. L., M. Pieja, and C. A. Floudas. 2003. A new class of hybrid global optimization algorithms for peptide structure prediction: alternating hybrids and application of met-enkephalin and melittin. *Biophys. J.* 84:869–882.
61. Klepeis, J. L., and C. A. Floudas. 2003. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* 85:2119–2146.
62. Klepeis, J. L., Y. N. Wei, M. H. Hecht, and C. A. Floudas. 2005. Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. *Proteins.* 58:560–570.
63. Maranas, C. D., and C. A. Floudas. 1997. Global optimization in generalized geometric programming. *Comput. Chem. Eng.* 21: 351–369.
64. Floudas, C. A., and P. M. Pardalos. 1995. State-of-the-art in global optimization. Computational methods and applications. *J. Glob. Optim.* 7:113.
65. Guntert, P. 2004. Automated NMR structure calculation with CYANA. *Methods Mol. Biol.* 278:353–378.
66. Guntert, P., C. Mumenthaler, and K. Wuthrich. 1997. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* 273:283–298.
67. Ponder, J. 1998. TINKER, Software Tools for Molecular Design. Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO.
68. Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.
69. Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins.* 65:712–725.
70. Koehl, P., and M. Levitt. 1999. De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.* 293:1161–1181.
71. Koehl, P., and M. Levitt. 1999. De novo protein design. II. Plasticity in sequence space. *J. Mol. Biol.* 293:1183–1193.
72. Melnyk, R. A., S. Kim, A. R. Curran, D. M. Engelman, J. U. Bowie, and C. M. Deber. 2004. The affinity of GXXXG motifs in transmembrane helix-helix interactions is modulated by long-range communication. *J. Biol. Chem.* 279:16591–16597.
73. Bowie, J. U. 1997. Helix packing in membrane proteins. *J. Mol. Biol.* 272:780–789.
74. Filipek, S., K. A. Krzysko, D. Fotiadis, Y. Liang, D. A. Saperstein, A. Engel, and K. Palczewski. 2004. A concept for G protein activation by G protein-coupled receptor dimers: the transducin/rhodopsin interface. *Photochem. Photobiol. Sci.* 3:628–638.