Identifying residue–residue clashes in protein hybrids by using a second-order mean-field approach

Gregory L. Moore and Costas D. Maranas*

Department of Chemical Engineering, Pennsylvania State University, 112 Fenske Laboratory, University Park, PA 16802

Communicated by Stephen J. Benkovic, Pennsylvania State University, University Park, PA, February 28, 2003 (received for review December 6, 2002)

In this article, a second-order mean-field-based approach is introduced for characterizing the complete set of residue–residue couplings consistent with a given protein structure. This information is subsequently used to classify protein hybrids with respect to their potential to be functional based on the presence/absence and severity of clashing residue–residue interactions. First, atomistic representations of both the native and denatured states are used to calculate rotamer–backbone, rotamer–intrinsic, and rotamer–rotamer conformational energies. Next, this complete conformational energy table is coupled with a second-order mean-field description to elucidate the probabilities of all possible rotamer–rotamer combinations in a minimum Helmholtz free-energy ensemble. Computational results for the dihydrofolate reductase family reveal correlation in substitution patterns between not only contacting but also distal second-order structural elements. Residue–residue clashes in hybrid proteins are quantified by contrasting the ensemble probabilities of protein hybrids against the ones of the original parental sequences. Good agreement with experimental data is demonstrated by superimposing these clashes against the functional crossover profiles of bidirectional incremental truncation libraries for *Escherichia coli* and human glycinamide ribonucleotide transformylases.

The use of DNA mutagenesis and/or recombination in the context of directed-evolution experiments has emerged as a leading strategy in protein engineering (1–3). However, the majority of generated protein hybrids have either substantially reduced or even completely lost functionalities. Therefore, the *a priori* classification of protein hybrids with respect to their potential to be functional is widely being recognized as an overarching challenge for many combinatorial protein-engineering efforts. In the past, the majority of successful combinatorial efforts involved the recombination of parental sequences sharing relatively high sequence identity (i.e., >70% at the DNA level). With the advent of a number of experimental protocols capable of recombining parental sequences with low sequence identity [e.g., ITCHY/SCRATCHY (4, 5), SHIPREC (6), GeneReassembly (7)], it has been observed that the fraction of functional hybrids in the combinatorial library decreases dramatically as the level of sequence identity shared in the parental set is reduced (5, 6). Given that most members of a protein family share pairwise sequence identities of <70%, this implies that a large portion of protein diversity may be left unexplored because of the scarcity of functional hybrids. This leads to the following dilemma: How can diversity generated by the recombination of low sequence identity parental sequences be explored effectively without severely curtailing the chances of success? To resolve this dilemma effectively, it is necessary to elucidate *a priori* what crossovers or crossover combinations are likely to lead to hybrids with preserved/improved functionality.

A number of hypotheses have been advanced to explain how crossovers affect the integrity of proteins. Monte Carlo simulations by Bogarad and Deem (8) suggested that the swapping of low-energy structures was least disruptive to protein structure, but delineating these structures has thus far not been straightforward. The SCHEMA algorithm (9) postulated structural disruption when a contacting residue pair in a hybrid does not match at least one of the parental proteins, and it was used to explain the crossover distributions found in a number of experiments. Although promising, this approach cannot differentiate between hybrids with different directionality (i.e., an A–B versus a B–A crossover), which have been shown to often have very different functional crossover profiles (5).

In our previous work, programs for estimating the frequency and location of crossovers in combinational DNA libraries were developed (5, 10, 11). In this article, the second-order mean-field identification of residue–residue clashes in protein hybrids (SIRCH) procedure for evaluating protein hybrids is introduced. Residue–residue clashes may arise because of a different directionality in the parental sequences with regard to a charged pair, residue sizes, or hydrogen bond (see Fig. 1), among other reasons. SIRCH consists of three steps. (*i*) Calculation of possible rotamer–backbone, rotamer–intrinsic, and rotamer–rotamer conformational energies (including van der Waals, electrostatic, and solvation contributions) by using atomistic representations of both the native and denatured states. (*ii*) Use of an extended, second-order mean-field description to elucidate the probabilities of all possible residue–residue combinations in a minimum Helmholtz free-energy ensemble. (*iii*) Systematic detection of clashes in potential hybrids through the evaluation of pairwise substitution patterns uncovered by the second-order mean-field description. A complete characterization of the entire collection of all possible residue–residue combinations complying with the protein family backbone coordinates is generated. This *in silico* protein family description augments the incomplete/coarse correlation statistics that can be gleaned from protein family sequence data. The SIRCH procedure is used to analyze pairwise substitution patterns in the dihydrofolate reductase (DHFR) enzyme family and to assess the result of the recombination of *Escherichia coli* and human glycinamide ribonucleotide (GAR) transformylases (5, 12, 13). Results demonstrate that experimentally determined functional crossover positions for the GAR transformylases are consistent with the predicted residue–residue clashes, capturing the effect of crossover directionality (i.e., an A–B versus a B–A crossover) observed in experimental crossover distributions.

Method

Conformational Energy Calculation. Conformational energy has been used widely (14–19) as a scoring function to query whether a particular hybrid protein will likely retain functionality or whether unfavorable energetic interactions and geometric clashes brought about by recombination will prevent the hybrid from even conforming to the backbone structure. Rotamer combinations (the term "rotamer" is used here to include side-chain conformers of all residue types) are used to describe hybrid protein conformations and designs. The protein family (and fold) of interest is represented by the backbone coordinates of a single representative structure. The coordinates of the backbone atoms along with any wild-type proline residues are locked throughout the calculation (neither Pro → X nor X → Pro

Abbreviations: SIRCH, second-order mean-field identification of residue–residue clashes in protein hybrids; DHFR, dihydrofolate reductase; GAR, glycinamide ribonucleotide.

*To whom correspondence should be addressed. E-mail: costas@psu.edu.

BIOCHEMISTRY

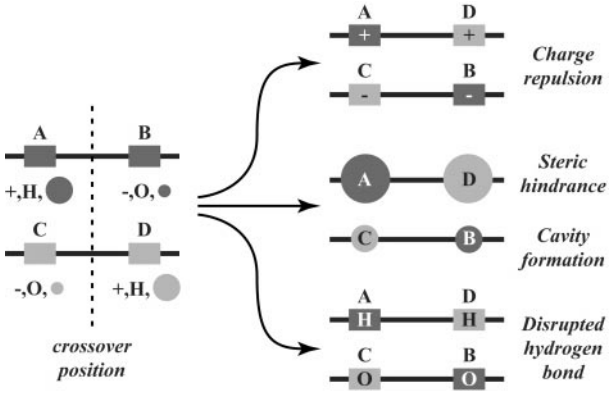

Fig. 1. Residue–residue clashes may arise in protein hybrids because of a different directionality in the parental sequences of a charged pair, residue sizes, or hydrogen bond. H, proton donor; O, proton acceptor.

mutations are permitted; also, cis/trans isomerization is not allowed).

The conformational energy of a rotamer combination in the native state is expressed as the sum of (*i*) rotamer–backbone energies, $e_i^{bb}(r)$, (*ii*) rotamer–intrinsic energies, $e_i^{int}(r)$, and (*iii*) rotamer–rotamer energies, $e_{ij}(rs)$. Here i and j refer to sequence positions, and r and s refer to rotamer choices at positions i and j, respectively. The total energy E of a specific combination of rotamers in the native state can be written as

$$E_{\text{combination}} = \sum_{i=1}^{N} e_i(r) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_{ij}(rs), \qquad [1]$$

where N represents the total number of residues in the protein, and $e_i(r) = e_i^{bb}(r) + e_i^{int}(r)$. The first two terms describe rotamer–backbone and rotamer–intrinsic interaction energies, while the third term describes rotamer–rotamer interaction energies. For every position, excluding the termini (1 and N), a backbone-dependent (i.e., on ϕ and ψ dihedral angles) set of rotamers is considered, in accordance with the library of Dunbrack and Cohen (20). For the termini, a backbone-independent rotamer library (20) is used. For each sequence position, the rotamer library (excluding proline rotamers) encompasses 320 different rotamer/residue combinations. Prior to the calculation of the rotamer–backbone and rotamer–intrinsic energies, rotamers are subjected to 50 steps of conjugate gradient minimization (18) by using CHARMM (21).

The CHARMM program is used along with version 22 of the all-atom parameters (22) to estimate conformational energies. Three contributions to conformational energy are considered: (*i*) van der Waals, (*ii*) electrostatics (including hydrogen bonds), and (*iii*) solvation. For both van der Waals and electrostatics, a cutoff distance of 14 Å is used without any scaling of the 1–4 interactions. A Coulombic potential is used with a constant dielectric constant ($\varepsilon = 8$) as suggested in ref. 18. Solvation energies are described as the sum of the solvation energies for the individual atoms in the rotamer. The solvation energy of each atom is assumed to be proportional to its accessible surface area as determined analytically by a 1.4-Å probe. The proportionality constants of Wesson and Eisenberg (23), developed specifically for use in CHARMM, are used to estimate solvation energies based on accessible surface areas. Rotamer–rotamer solvation energies are estimated by using the method of Street and Mayo (24), in which the difference in solvation energy due to the overlap of two isolated side chains is scaled down by 50% to prevent overcounting.

The three contributions to conformational energy are used without any empirical balancing. However, comparison of rotamers of different types can be misleading without the use of a reference energy (18). For instance, without consideration of a reference energy, arginine residues are highly favored over other types because of their high solubility and large size. Therefore, the establishment of a reference state for each of the different residue types is necessary for providing a consistent basis of comparison. We use the "expanded" state of Elcock (25) to represent the denatured-state ensemble, allowing the calculation of standardized rotamer energy differences $\delta e_i(r)$ and standardized rotamer–rotamer energy differences $\delta e_{ij}(rs)$. This representation of the denatured state has two advantages over dipeptide/tripeptide systems. First, the number and type of atoms remain constant, and second, the topology of the protein fold is retained such that atoms that are in close proximity in the native state remain relatively close to each other in the denatured state. This procedure is described in detail in *Supporting Text*, which is published as supporting information on the PNAS web site, www.pnas.org. A depiction of the expanded state is also found in Fig. 4, which is published as supporting information on the PNAS web site. The standardized conformational energy ΔE for a specific rotamer combination can then be written as

$$\Delta E_{\text{combination}} = \sum_{i=1}^{N} \delta e_i(r) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \delta e_{ij}(rs). \qquad [2]$$

Prior to the calculation of rotamer–rotamer conformational energies, rotamers are screened out of the library if $\delta e_i(r)$ is >50 kcal/mol or they are not among the 10 lowest energy choices for a particular residue type (19). Typically, ≈100–120 rotamers are retained for each sequence position, encompassing all residue choices considered.

Ensemble of Rotamer/Residue States. The objective of this study is to determine whether a residue–residue pair brought about by recombination and/or mutation is structurally favorable or unfavorable. This necessitates the establishment of the proper tradeoff between structural fitness (energy) and sequence/conformational variation (entropy) characteristic of protein families. To this end, a statistical mechanics description of the residue/rotamer space of states (ensemble) is adopted. An ensemble of states is defined as the collection of all possible rotamer and residue combinations. The membership probabilities $\mathcal{P}$ of each state are found by equilibrating the ensemble. The expressions for the total energy and entropy of the ensemble, containing not only different rotamer choices but also different residue choices for each sequence position, are functions of the respective state probabilities $\mathcal{P}$, as shown next.

$$U_{\text{ensemble}} = \sum_{\substack{\text{all rotamer} \\ \text{combinations}}} \mathcal{P}_{\text{combination}} \Delta E_{\text{combination}} \qquad [3]$$

$$S_{\text{ensemble}} = -R \sum_{\substack{\text{all rotamer} \\ \text{combinations}}} \mathcal{P}_{\text{combination}} \ln \mathcal{P}_{\text{combination}} \qquad [4]$$

Assuming a canonical ensemble (a closed system with constant temperature T), the state probabilities are determined at equilibrium by minimizing the Helmholtz free energy $A_{\text{ensemble}} = U_{\text{ensemble}} - TS_{\text{ensemble}}$. The use of the Helmholtz free energy allows the systematic exploration of tradeoffs between conformational energy and entropy. However, the direct solution of this problem is intractable, because the number of possible rotamer/residue choices is prohibitively large. For example, a 200-residue protein with 120 rotamer choices for each position gives rise to $120^{200} \approx 10^{416}$ possible rotamer combinations. Mean-field approximations are used to restore tractability to the ensemble-equilibration problem.

First-Order Mean-Field Approximation. Earlier mean-field approximations to the Helmholtz free energy (14, 26, 27), referred to herein as first-order, were based on the assumption that the probability $\mathcal{P}$

of a specific rotamer combination can be approximated as the product of individual rotamer site probabilities $p_i(r)$ of each sequence position i. This implies that the site probabilities at each position are assumed to vary independently from one another.

$$\mathcal{P}^{(1)}_{\text{combination}} = \prod_{i=1}^{N} p_i(r) \qquad [5]$$

This simplification substantially reduces the number of state probabilities required to describe the ensemble (e.g., from 10^{416} to $200 \cdot 120 = 24{,}000$ for a 200-residue protein). Substituting the first-order approximation (Eq. **5**) into the expressions for the energy and entropy of a rotamer sequence (Eqs. **3** and **4**) leads to the following expressions for the first-order mean-field energy $U^{(1)}$ and entropy $S^{(1)}$ of the ensemble,

$$U^{(1)} = \sum_{i=1}^{N} \sum_{r\in\mathcal{R}_i} p_i(r)\delta e_i(r) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{r\in\mathcal{R}_i} \sum_{s\in\mathcal{R}_j} p_i(r)p_j(s)\delta e_{ij}(rs) \qquad [6]$$

$$S^{(1)} = -R \sum_{i=1}^{N} \sum_{r\in\mathcal{R}_i} p_i(r) \log p_i(r), \qquad [7]$$

where $\mathcal{R}_i$ and $\mathcal{R}_j$ represent the set of rotamer choices available at positions i and j, respectively. Minimization of the first-order mean-field free energy $A^{(1)} = U^{(1)} - TS^{(1)}$, subject to the condition that the site probabilities sum up to one ($\Sigma_{r\in\mathcal{R}_i} p_i(r) = 1$), yields

$$p_i(r) = \frac{\exp(-\bar{\delta}e_i(r)/RT)}{\sum_{r'\in\mathcal{R}_i} \exp(-\bar{\delta}e_i(r')/RT)}, \quad \forall i, r \in \mathcal{R}_i, \qquad [8]$$

where

$$\bar{\delta}e_i(r) = \delta e_i(r) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{s\in\mathcal{R}_j} p_j(s)\delta e_{ij}(rs), \quad \forall i, r \in \mathcal{R}_i. \qquad [9]$$

The mean-field energy $\bar{\delta}e_i(r)$ can be interpreted as the energy of rotamer r placed at sequence position i plus the average interaction energy that it experiences from other rotamer choices s at other positions j in the ensemble. As shown in Eq. **8**, the site probabilities are Boltzmann-distributed with respect to their mean-field energies. Typical solution procedures involve uniform initialization of the rotamer probabilities and iterative calculation of the mean-field energies (Eq. **9**) and site probabilities (Eq. **8**) until self-consistency is achieved (26–29). Koehl and Delarue (26) and Lee (27) used a first-order mean-field approach for estimating the conformational entropy of side chains and positioning them. Voigt *et al.* (14) and Saven and coworkers (19, 30) extended the ensemble to include both residue and rotamer choices to investigate the fitness of single residue substitutions in mutagenesis experiments.

A key limitation of the first-order mean-field approximation is that it cannot capture whether and/or how the substitution patterns at two sequence positions i and j are related. Therefore, no information can be gleaned as to how a site probability distribution at one position is influenced by placing a specific rotamer at another position (i.e., conditional probability). However, this is exactly the type of information needed to evaluate the impact of bringing together two new sets of residues in hybrids generated by recombination. To overcome these limitations, a second-order mean-field approximation to the Helmholtz free energy is developed that allows for the explicit consideration of rotamer–rotamer joint probabilities.

Second-Order Mean-Field Approximation. A second-order approximation is proposed that can track joint probabilities explicitly, represented by $P_{ij}(rs)$. The Bethe approximation (31) is used to estimate the ensemble probability $\mathcal{P}$ as the product of all joint probabilities, appropriately scaled to avoid double counting.

$$\mathcal{P}^{(2)}_{\text{combination}} = \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} P_{ij}(rs) \Bigg/ \prod_{i=1}^{N} p_i(r)^{N-2} \qquad [10]$$

The Bethe approximation was developed originally to assess the entropy within metallic superlattices (31, 32), but in recent years it has been applied in the field of computer vision (33) and has been shown to be analogous to the use of belief propagation methods (34) in resolving Bayesian causal networks (35).

Substituting the second-order mean-field approximation (Eq. **10**) into the equations for ensemble energy (Eq. **3**) and entropy (Eq. **4**) leads to the following expressions.

$$U^{(2)} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{r\in\mathcal{R}_i} \sum_{s\in\mathcal{R}_j} P_{ij}(rs)(\delta e_{ij}(rs) + \delta e_i(r) + \delta e_j(s)) - (N-2) \sum_{i=1}^{N} \sum_{r\in\mathcal{R}_i} p_i(r)\delta e_i(r) \qquad [11]$$

$$S^{(2)} = -R \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{r\in\mathcal{R}_i} \sum_{s\in\mathcal{R}_j} P_{ij}(rs) \ln P_{ij}(rs) - (N-2) \sum_{i=1}^{N} \sum_{r\in\mathcal{R}_i} p_i(r) \ln p_i(r) \right] \qquad [12]$$

As described earlier, the minimization of the ensemble free energy for the first-order mean-field approximation can readily be converted into a recursive relation resolved through direct substitution. Such a conversion for a second-order mean-field approximation is much more elusive. To accomplish this, a set of variable transformations is needed. First, the energy expression can be written in a form analogous to that of the entropy by substituting $\phi_i(r) = \exp(-\delta e_i(r)/RT)$ and $\psi_{ij}(rs) = \exp(-\delta e_{ij}(rs)/RT)$ into the expressions for the second-order energy and entropy (Eqs. **11** and **12**). By combining the resulting expressions via $A^{(2)} = U^{(2)} - TS^{(2)}$, the following expression for the Bethe free energy (scaled by RT) is derived.

$$\frac{A^{(2)}}{RT} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{r\in\mathcal{R}_i} \sum_{s\in\mathcal{R}_j} P_{ij}(rs)[\ln P_{ij}(rs) - \ln(\psi_{ij}(rs)\phi_i(r)\phi_j(s))] - (N-2) \sum_{i=1}^{N} \sum_{r\in\mathcal{R}_i} p_i(r)(\ln p_i(r) - \ln \phi_i(r)) \qquad [13]$$

The joint probabilities $P_{ij}(rs)$ are then equilibrated in the ensemble by minimizing $A^{(2)}/RT$, subject to

$$\sum_{r\in\mathcal{R}_i} p_i(r) = 1, \quad \forall i \qquad [14]$$

$$\sum_{r\in\mathcal{R}_i} \sum_{s\in\mathcal{R}_j} P_{ij}(rs) = 1, \quad \forall i, j > i \qquad [15]$$

$$\sum_{s\in\mathcal{R}_j} P_{ij}(rs) = p_i(r), \quad \forall i, j \neq i, r \in \mathcal{R}_i \qquad [16]$$

Eqs. **14** and **15** ensure that both the site and joint probability choices sum to one for a given position or pair of positions, respectively, whereas Eq. **16** ensures consistency between joint probabilities and respective site probabilities. The dimensionality of the resulting nonlinear optimization problem is too high to allow for direct numerical solution. For example, for a 200-residue protein, $>10^8$ probability variables are present. To remedy this, we use the method of Lagrangean multipliers for converting a constrained nonlinear optimization problem into a system of nonlinear algebraic equations. The Lagrangean function $\mathcal{L}$ is formed by augmenting the original function $A^{(2)}/RT$ by adding all three constraints to the objective function with multipliers γ_i, Γ_{ij}, and $\lambda_{ji}(r)$, respectively.

$$\begin{aligned}\mathcal{L} = \frac{A^{(2)}}{RT} &+ \sum_{i=1}^{N} \gamma_i \left(1 - \sum_{s\in\mathcal{R}_j} p_i(r)\right) \\ &+ \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \Gamma_{ij}\left(1 - \sum_{r\in\mathcal{R}_i}\sum_{s\in\mathcal{R}_j} P_{ij}(rs)\right) \\ &+ \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{r\in\mathcal{R}_i} \lambda_{ji}(r)\left(p_i(r) - \sum_{s\in\mathcal{R}_j} P_{ij}(rs)\right) \qquad [17]\end{aligned}$$

Minima of $\mathcal{L}$ are located at points where derivatives with respect to each of the variables (i.e., rotamer probabilities and multipliers) are equal to zero. Setting $\partial\mathcal{L}/\partial p_i(r) = 0$ yields

$$p_i(r) = z_i\phi_i(r)\exp\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{\lambda_{ji}(r)}{N-2}\right), \quad \forall i, r\in\mathcal{R}_i, \qquad [18]$$

where z_i is chosen to normalize $p_i(r)$ (Eq. **14**). Similarly, $\partial\mathcal{L}/\partial P_{ij}(rs) = 0$ provides

$$P_{ij}(rs) = Z_{ij}\psi_{ij}(rs)\phi_i(r)\phi_j(s)\exp(\lambda_{ji}(r) + \lambda_{ij}(s)), \quad \forall i, j > i, r\in\mathcal{R}_i, s\in\mathcal{R}_j, \qquad [19]$$

where Z_{ij} enforces the normalization of $P_{ij}(rs)$ (Eq. **15**).

Note that when the derivatives of $\mathcal{L}$ with respect to the multipliers are set to zero, the original three constraints (Eqs. **14–16**) are recovered. The set of five nonlinear equations (Eqs. **14–16**, **18**, and **19**) is recast further by substituting message variables $m_{ij}(s)$ for multipliers $\lambda_{ij}(s)$.

$$\lambda_{ij}(s) = \ln\prod_{\substack{k=1\\ k\neq i,j}}^{N} m_{kj}(s), \quad \forall i, j\neq i, s\in\mathcal{R}_j \qquad [20]$$

This variable substitution is motivated by methods used to resolve Bayesian networks by belief propagation (34). The message variables $m_{ij}(s)$ describe how the set of rotamer choices at position i interacts with the choice of rotamer s at position j, providing the following expression for $p_i(r)$.

$$p_i(r) = z_i\phi_i(r)\prod_{\substack{j=1\\ j\neq i}}^{N} m_{ji}(r), \quad \forall i, r\in\mathcal{R}_i \qquad [21]$$

An expression for $P_{ij}(rs)$ is derived in a similar fashion.

$$P_{ij}(rs) = Z_{ij}\psi_{ij}(rs)\phi_i(r)\phi_j(s)\left(\prod_{\substack{k=1\\ k\neq i,j}}^{N} m_{ki}(r)m_{kj}(s)\right), \quad \forall i, j > i, r\in\mathcal{R}_i, s\in\mathcal{R}_j \qquad [22]$$

Eqs. **21** and **22** then are combined via Eq. **16** to derive a recursion of reduced dimensionality, also known as belief propagation, containing only the message variables.

$$m_{ij}(s) = \sum_{r\in\mathcal{R}_i}\phi_i(r)\psi_{ij}(rs)\left(\prod_{\substack{k=1\\ k\neq i,j}}^{N} m_{ki}(r)\right), \quad \forall i, j\neq i, s\in\mathcal{R}_j \qquad [23]$$

Three factors are considered in the belief propagation recursion: (*i*) how rotamers at position i fit with rotamer s at position j ($\psi_{ij}(rs)$); (*ii*) how rotamers at position i fit the backbone ($\Sigma_r\,\phi_i(r)$); and (*iii*) how other positions k interact with rotamers at position i ($\Pi_k\,m_{ki}(r)$). Self-consistent resolution of this recursion yields values for the message variables, which then are substituted into Eqs. **21** and **22** to calculate the site and joint probabilities. Site and joint probabilities for specific residues a and residue pairs a, b are examined by aggregating the corresponding rotamer probabilities (where $\mathcal{R}_i^a$ represents the set of rotamers of residue type a available at position i).

$$p_i(a) = \sum_{r\in\mathcal{R}_i^a} p_i(r); \qquad P_{ij}(ab) = \sum_{r\in\mathcal{R}_i^a}\sum_{s\in\mathcal{R}_j^b} P_{ij}(rs) \qquad [24]$$

A flowchart summarizing the steps of the complete computational procedure is shown in Fig. 5, which is published as supporting information on the PNAS web site. With the second-order mean-field approximation in place, the correct temperature of the ensemble is estimated by matching the entropy of the natural Pfam (36) protein family to the entropy of the ensemble (see *Supporting Text* and Fig. 6, which is published as supporting information on the PNAS web site, for details).

Substitution Dependency D_{ij}. The identified site and joint ensemble probabilities are used to determine the tolerance of the protein structure, or lack thereof, for different residue combinations. Residue pairs that are favorable or unfavorable can be identified by examining the probability ratio $\alpha_{ij}(ab)$ that quantifies the departure of the joint probabilities from the independent substitution assumption. Specifically,

$$\alpha_{ij}(ab) = \frac{P_{ij}(ab)}{p_i(a)p_j(b)}\begin{cases} > 1, a \text{ and } b \text{ are favored at } i, j \\ < 1, a \text{ and } b \text{ are unfavored at } i, j \\ = 1, \text{no preference.}\end{cases} \qquad [25]$$

The standard deviation of $\alpha_{ij}(ab)$ over all residue combinations provides a quantitative metric for the substitution dependency D_{ij}:

$$D_{ij} = \left[\sum_{a=1}^{20}\sum_{b=1}^{20} P_{ij}(ab)(\log_2\alpha_{ij}(ab) - \mu_{ij})^2\right]^{1/2} \qquad [26]$$

$$\text{where } \mu_{ij} = \sum_{a=1}^{20}\sum_{b=1}^{20} P_{ij}(ab)\log_2\alpha_{ij}(ab).$$

A zero value for the substitution dependency D_{ij} implies that residue positions i and j have independent substitution patterns. Nonzero (positive) values for D_{ij} signify correlation in the substitution patterns. The larger the value of D_{ij}, the stronger the correlation is between positions i and j. The substitution-dependency metric D_{ij} along with the probability ratios $\alpha_{ij}(ab)$ can be used not only for elucidating substitution correlation between two residue positions but also for querying whether residue pairs in a protein hybrid comply or clash with the family protein structure in comparison to the parental sequences.

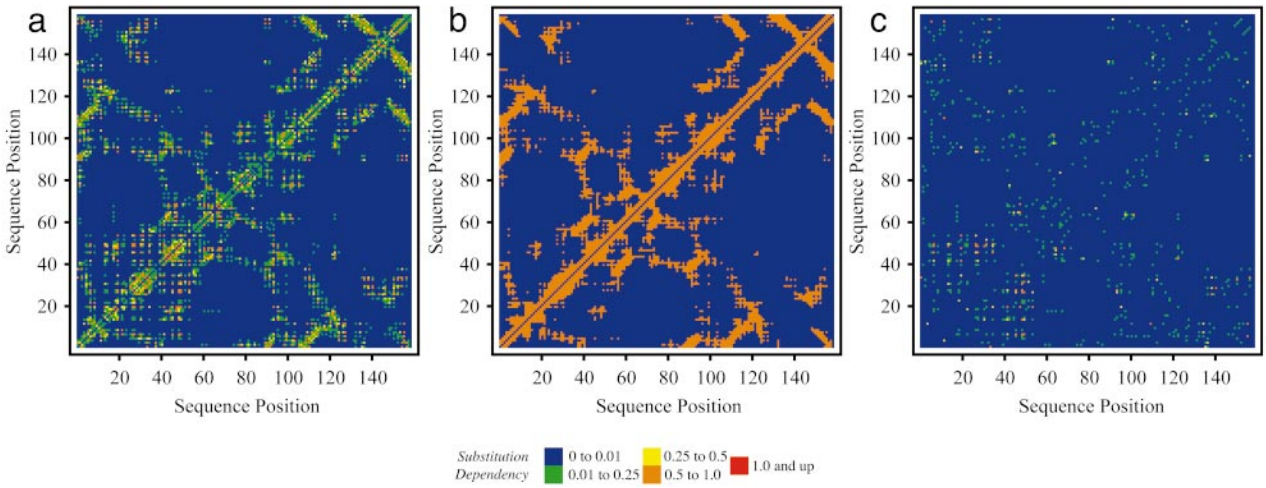

Fig. 2. (*a*) Map of substitution dependency for *E. coli* DHFR, closed M20 (1rx2). (*b*) Contact map (<8 Å) for 1rx2. Orange denotes contacting residue pairs. (*c*) Map of substitution dependency after removing contacting residue pairs depicted in *a* for 1rx2.

Correlation in the Substitution Patterns of the DHFR Protein Family

The well studied DHFR protein family is first addressed to examine whether well known correlated substitution patterns can be revealed by SIRCH. The substitution dependencies D_{ij} based on four different DHFR crystals [i.e., *E. coli*: PDB ID code 1rx2, M20 closed (37), PDB ID code 1rx5, M20 occluded (37), and PDB ID code 1ra9, M20 open (37); and *Lactobacillus casei*: PDB ID code 3dfr, M20 closed (38)] downloaded from the Protein Data Bank (39) are calculated. The first three crystals are snapshots of important steps in the *E. coli* DHFR catalytic cycle (37), whereas the fourth is a non-*E. coli* DHFR. Fig. 7 (which is published as supporting information on the PNAS web site) depicts the substitution dependency plots for the four structures. The plots are almost identical, demonstrating that the choice of crystal does not alter the results substantially. The only significant difference is between the results for the open M20 structure (1ra9) and the two closed structures (1rx2 and 3dfr). Specifically, for the closed structures, residues 25–50 exhibit a more pronounced substitution dependency. This is consistent with the fact that in the closed conformation residues 25–50 are approached by the M20 loop and other connecting residues.

In the residue–residue substitution-dependency plot for 1rx2 (Fig. 2*a*), blue implies no correlation, whereas green, yellow, orange, and red depict residue pairs with increased levels of correlation in substitution patterns. Interestingly, strong correlation between the contacting M20 and FG loops (i.e., residues 7–24 and 116–132, respectively) as well as between the end of the M20 loop (residues 20–25) and the GH loop (residues 142–150) is predicted correctly. Quite remarkably, strong correlation between the M20/Hinge region (20–38) with both the region from residues 45–50 and the region from residues 93–97 is also elucidated even though these domains are not contacting (distance >8 Å), alluding to the fact that correlation information seems to be propagated through a network of interacting residues. The ability of the method to capture distal correlations in substitution patterns is shown more clearly in Fig. 2 *b* and *c*, in which the substitution-dependency density plot is contrasted against the set of contacting residues. It appears that important correlation information between residue pairs is encoded within D_{ij} that does not necessarily require them to be contacting. Another important observation involves a comparison of the residue pairs that exhibit correlated motion (in the same direction) based on the molecular dynamics study of Radkiewicz and Brooks (40), and the substitution-dependency plot (see Fig. 8, which is published as supporting information on the PNAS web site). The strong similarity between the two alludes that residues that "move" in the same direction must also be substituted in a coordinated manner.

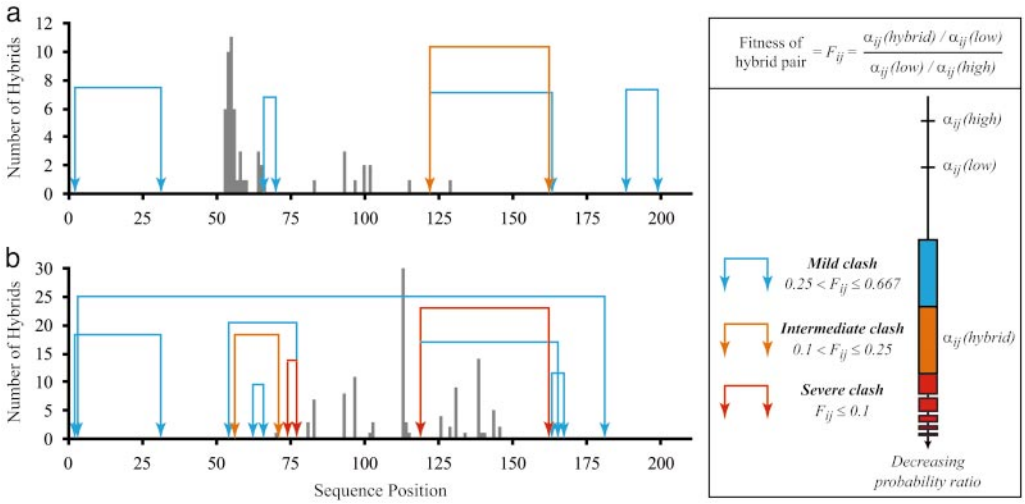

Fig. 3. Clashing residue pairs in human/*E. coli* (*a*) and *E. coli*/human (*b*) hybrids. Clashes are classified as mild, intermediate, or severe based on the fitness metric F_{ij}, which is calculated by comparing the probability ratio of the hybrid residue pair α_{ij}(hybrid) to th probability ratios of the parental sequences α_{ij}(low), α_{ij}(high), where low refers to the parental sequence with the lower α_{ij}, and high refers to the higher-valued one. Vertical bars indicate positions where functional crossovers have been found in incremental truncation experiments (5, 12, 13).

Next, the *a priori* classification of crossovers with respect to their functionality through SIRCH is addressed. This is accomplished by contrasting the experimental results for the *E. coli* and human GAR-transformylase system with the model predictions.

In Silico GAR-Transformylase Hybrid Prescreening

By using the structure of *E. coli* GAR transformylase [PDB ID code 1gar (41)] as a reference, SIRCH is used to characterize all single-crossover hybrids between *E. coli* and human versions of GAR transformylase (protein sequence identity of 45%). The locations of all functional crossovers in bidirectional hybrids generated through incremental truncation (5, 12, 13) are depicted as vertical bars in Fig. 3. The incremental truncation window is between residues 50 and 150. Clearly, functional crossovers are distributed quite differently depending on the directionality of the incremental truncation library (compare Fig. 3 *a* and *b*).

Residue–residue clashes predicted for single-crossover hybrids are shown pictorially as arcs of different colors linking the corresponding residues (see Fig. 3). These clashes are present only in hybrids with a crossover positioned between the two residues (i.e., cutting the arc). The severity of the clash is quantified by contrasting the hybrid residue pair probability ratio against the probability ratios corresponding to the two parental (wild-type) sequences (i.e., *E. coli* and human). By using the parental residue pairs as a baseline, the comparison only reveals clashes generated in the hybrid that are absent in the parental sequences. Blue arcs signify a relatively small difference in probability ratio between the hybrid and the parental sequences, whereas orange and red arcs denote clashes of increasing intensity based on the hybrid/parental sequence probability ratio difference. For the human/*E. coli* library (Fig. 3*a*), a large cluster of functional crossovers is present at the beginning of the recombination range, followed by an abrupt end at position 66. Remarkably, position 66 is the location of the first residue for the first clash in the recombination window. Past the first clashing pair, a few functional crossovers are present that again disappear after encountering a pair of nested clashes. Unlike the human/*E. coli* library, no functional crossovers are present at the beginning of the recombination range for the *E. coli*/human library (Fig. 3*b*), which is consistent with the numerous clashes found within the range of 54–77. A large number of functional crossovers (81–115) violates only a mild clash, whereas the group between positions 125 and 150 is inconsistent with a severe clash between residues 119 and 162. Molecular modeling for these two positions reveals a steric hindrance between histidine and valine that cannot be relieved without substantial backbone movement. In this case, it seems that this movement did not affect catalytic activity or binding affinity, pointing at some of the limitations of mean-field-based approximation techniques. Overall, SIRCH seems to be quite successful, although not perfect, at classifying crossovers in terms of their potential to yield functional hybrids. More importantly, by identifying a relatively small set of clashing residue combinations, SIRCH provides valuable information for designing strategies based on site-directed mutagenesis for relieving these clashes.

Summary

In this article, a second-order mean-field approach was described for the complete description of the entire residue substitution space of a protein family. The procedure was implemented in the SIRCH program (see fenske.che.psu.edu/faculty/cmaranas) for identifying and quantifying the severity of residue–residue clashes in protein hybrids. This information can then be used upstream or downstream to suggest site-directed mutagenesis strategies for either (*i*) the parental sequences or (*ii*) hybrids with residual functionalities that will lead to the reduction or elimination of clashes in the protein combinatorial library. Note that the obtained results were largely insensitive to the starting protein crystal and that a strong correlation between residue substitution-dependency patterns and residue motions in the crystal was observed.

Computational results uncovered correlated substitution patterns for the DHFR family not only between contacting but also between widely separated domains, alluding to the propagation of residue substitution correlation information through a network of interacting residues (42). In addition, the distribution of functional crossovers for the incremental truncation libraries (5, 12, 13) of *E. coli*/human GAR and human/*E. coli* GAR transformylases was in very good agreement with the residue–residue clashes revealed by SIRCH. These results are currently being used to identify site-directed mutagenesis strategies for ratcheting up the functionality of barely active hybrids. Thus far, the only information gleaned from the sequence data of protein families (39) involved setting the entropy of the computationally equilibrated ensemble. Nevertheless, additional restrictions can be imported into the ensemble by appending appropriate equality or even inequality constraints. These constraints may, for example, fix the consensus active-site residues, restrict the fraction of charged residues present in the library, or establish hydrophobic/polar patterning requirements.

We thank Professor Stephen Benkovic, Dr. Alexander Horswill, and Dr. Anshuman Gupta for helpful discussions and the reviewers for useful suggestions. Financial support from National Science Foundation Award BES0120277 and hardware support by the IBM-SUR program are gratefully acknowledged.

1. Petrounia, I. P. & Arnold, F. H. (2000) *Curr. Opin. Biotechnol.* **11,** 325–330.
2. Brakmann, S. (2001) *Chembiochem* **2,** 865–871.
3. Schmidt-Dannert, C. (2001) *Biochemistry* **40,** 13125–13136.
4. Ostermeier, M., Nixon, A. E. & Benkovic, S. J. (1999) *Bioorg. Med. Chem.* **7,** 2139–2144.
5. Lutz, S., Ostermeier, M., Moore, G. L., Maranas, C. D. & Benkovic, S. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 11248–11253.
6. Sieber, V., Martinez, C. A. & Arnold, F. A. (2001) *Nat. Biotechnol.* **19,** 456–460.
7. Short, J. M. (1999) U.S. Patent 5,965,408.
8. Bogarad, L. D. & Deem, M. W. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2591–2595.
9. Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L. & Arnold, F. H. (2002) *Nat. Struct. Biol.* **9,** 553–558.
10. Moore, G. L., Maranas, C. D., Lutz, S. & Benkovic, S. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3226–3231.
11. Moore, G. L. & Maranas, C. D. (2000) *J. Theor. Biol.* **205,** 483–503.
12. Ostermeier, M., Shim, J. H. & Benkovic, S. J. (1999) *Nat. Biotechnol.* **17,** 1205–1209.
13. Lutz, S., Ostermeier, M. & Benkovic, S. J. (2001) *Nucleic Acids Res.* **29,** e16.
14. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z.-G. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3778–3783.
15. Dahiyat, B. I. & Mayo, S. L. (1997) *Science* **278,** 82–87.
16. Koehl, P. & Levitt, M. (1999) *J. Mol. Biol.* **293,** 1161–1181.
17. Raha, K., Wollacott, A. M., Italia, M. J. & Desjarlais, J. R. (2000) *Protein Sci.* **9,** 1106–1119.
18. Wernisch, L., Hery, S. & Wodak, S. J. (2000) *J. Mol. Biol.* **301,** 713–736.
19. Kono, H. & Saven, J. G. (2001) *J. Mol. Biol.* **306,** 607–628.
20. Dunbrack, R. L., Jr., & Cohen, F. E. (1997) *Protein Sci.* **6,** 1661–1681.
21. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
22. MacKerell, A. D., Jr., Bashford, D., Bellott, M., Dunbrack, R. L., Jr., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., *et al.* (1998) *J. Phys. Chem. B* **102,** 3586–3616.
23. Wesson, L. & Eisenberg, D. (1992) *Protein Sci.* **1,** 227–235.
24. Street, A. G. & Mayo, S. L. (1998) *Folding Des.* **3,** 253–258.
25. Elcock, A. H. (1999) *J. Mol. Biol.* **294,** 1051–1062.
26. Koehl, P. & Delarue, M. (1994) *J. Mol. Biol.* **239,** 249–275.
27. Lee, C. (1994) *J. Mol. Biol.* **236,** 918–939.
28. Koehl, P. & Delarue, M. (1995) *Nat. Struct. Biol.* **2,** 163–170.
29. Vasquez, M. (1995) *Biopolymers* **36,** 53–70.
30. Zou, J. & Saven, J. G. (2000) *J. Mol. Biol.* **296,** 281–294.
31. Bethe, H. A. (1935) *Proc. R. Soc. London Ser. A* **150,** 552–575.
32. Pathria, R. K. (1996) *Statistical Mechanics* (Butterworth–Heinemann, Boston).
33. Freeman, W. T., Pasztor, E. C. & Carmichael, O. T. (2000) *Int. J. Comput. Vis.* **40,** 25–47.
34. Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Kaufmann, San Francisco).
35. Yedidia, J. S. (2001) in *Advanced Mean Field Methods: Theory and Practice*, eds. Opper, M. & Saad, D. (MIT Press, Cambridge, MA).
36. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002) *Nucleic Acids Res.* **20,** 276–280.
37. Sawaya, M. R. & Kraut, J. (1997) *Biochemistry* **36,** 586–603.
38. Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. & Kraut, J. (1982) *J. Biol. Chem.* **257,** 13650–13662.
39. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
40. Radkiewicz, J. L. & Brooks, C. L., III (2000) *J. Am. Chem. Soc.* **122,** 225–231.
41. Klein, C., Chen, P., Arevalo, J. H., Stura, E. A., Marolewski, A., Warren, M. S., Benkovic, S. J. & Wilson, I. A. (1995) *J. Mol. Biol.* **249,** 153–175.
42. Agarwal, P. K., Billeter, S. R., Rajagopalan, P. T., Benkovic, S. J. & Hammes-Schiffer, S. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2794–2799.