*J. Biochem.* **140**, 543–552 (2006)
doi:10.1093/jb/mvj184

# A Monte Carlo Sampling Method of Amino Acid Sequences Adaptable to Given Main-Chain Atoms in the Proteins

**Koji Ogata[1], Kenji Soejima[2,*] and Junichi Higo[3]**

*[1]Centre for Computational Biology, The Hospital for Sick Children, 555 University Avenue, Toronot Ontario M5G 1X8, Canada; [2]Research Department 1, The Chemo-Sero-Therapeutic Research Institute, Kyokushikawabe, Kikuchi, Kumamoto 869-1298; and [3]Laboratory of Bioinformatics, School of Life Science, Tokyo University of Pharmacy and Life Science, 1432-1 Horinouchi, Hachioji, Tokyo 192-0392*

Received August 1, 2006; accepted August 26, 2006

**We have developed a computational method of protein design to detect amino acid sequences that are adaptable to given main-chain coordinates of a protein. In this method, the selection of amino acid types employs a Metropolis Monte Carlo method with a scoring function in conjunction with the approximation of free energies computed from 3D structures. To compute the scoring function, a side-chain prediction using another Metropolis Monte Carlo method was performed to select structurally suitable side-chain conformations from a side-chain library. In total, two layers of Monte Carlo procedures were performed, first to select amino acid types (1st layer Monte Carlo) and then to predict side-chain conformations (2nd layers Monte Carlo). We applied this method to sequence design for the entire sequence on the SH3 domain, Protein G, and BPTI. The predicted sequences were similar to those of the wild-type proteins. We compared the results of the predictions with and without the 2nd layer Monte Carlo method. The results revealed that the two-layer Monte Carlo method produced better sequence similarity to the wild-type proteins than the one-layer method. Finally, we applied this method to neuraminidase of influenza virus. The results were consistent with the sequences identified from the isolated viruses.**

**Key words: free energy, Monte Carlo method, protein design, sequence prediction, side-chain prediction.**

Prediction of the tertiary structure of a protein from a given amino acid sequence (forward folding) remains difficult because it requires finding the optimal sequences from a huge number of amino acid combinations and the corresponding conformational space of side-chains. To predict the folding of even a small protein, a long computational time using hundreds to thousands of parallelized computers is required (*1*, *2*).

In contrast, the inverse-folding approach, known as protein design, has been applied to a wide range of biological problems with successful results (*3*–*26*). This approach can predict the sequences adaptable to a given backbone structure that are optimized using a scoring function defined by various physicochemical and statistical parameters. The predicted sequences will indicate the stable proteins, and thus this method can be applied to obtaining stable mutant proteins.

In the protein design method, a large number of sequences and wide side-chain conformational space must be examined to obtain the optimal sequence and the best side-chain conformation (*27*), respectively. To achieve these examinations, an accurate scoring function to select the sequence (*28*, *29*) and an accurate prediction method of side-chain conformations are required. Appropriate combination of the scoring function and the side-chain prediction method will increase the accuracy of the determined sequences.

A number of side-chain prediction methods have been developed and reported to yield accurate results (*30*, *31*). The methods producing high efficiency searches of the optimal side-chain conformations include the Monte Carlo method (*16*, *32*), genetic algorithm (*10*, *33*), and dead-end elimination (*6*, *34*).

The scoring function should be based on an evaluation of the difference in thermodynamic stability between two mutant proteins. This theory is similar to that of a free-energy perturbation method, in which the free-energy difference resulting from a single amino acid mutation, $\Delta\Delta G$, is computed through a molecular dynamics simulation (*35*). The perturbation method, however, is time-consuming and difficult to perform for multiple mutations, making this method ineffective for protein design. Scoring functions using physicochemical parameters (*34*), statistical parameters (*36*), or combinations of these parameters (*33*) are frequently applied to protein design, although these functions only give a rough approximation of $\Delta\Delta G$. These scoring functions, however, result in a considerably shorter computational time than more accurate methods (*i.e.*, the free-energy perturbation method). Rapid computation makes possible the examination of a large number of side-chain conformations. Therefore, we can expect to obtain energetically and structurally acceptable amino acid sequences for a given main-chain using this accurate scoring function.

*To whom correspondence should be addressed. Tel: +81-968-37-3100, Fax: +81-968-37-3616, E-mail: soejima@kaketsuken.or.jp

 

In this study, we have developed a structure-based protein design method using a scoring function that approximates $\Delta\Delta G$. In this method, the optimal side-chain conformation at a backbone position is selected from a large side-chain library using a Metropolis Monte Carlo method (2nd layer Monte Carlo procedure). During this procedure, the free energy, $\Delta G$, is computed by counting the various side-chain conformations sampled. The amino acid type at that position is then selected with a second Monte Carlo method (1st layer Monte Carlo procedure). From these two Monte Carlo procedures, the scoring function corresponding to $\Delta\Delta G$ is computed, allowing sequences adaptable to the backbone to be obtained. We assessed the accuracy of this method by the application to three systems, SH3, Protein G, and BPTI. We demonstrated that the main-chain conformations of these proteins have a structural space that acceptable using a variety of amino acid sequences in which the specific amino acids differed from those in the wild-type protein, but had similar properties.

We then predicted the amino acid types in the influenza virus neuraminidase (NA) molecule. Without treatment, infection with influenza virus causes significant mortality worldwide. The influenza virus has two major glycoproteins, hemagglutinin (HA) and NA, on the viral surface. This virus has a remarkable capacity to escape from the host immune system by changing the antigenicity of its surface proteins. This variability is especially potent for the HA molecule, which has both receptor binding and fusion activities that initiate infection of the host cell. The NA molecule plays an important role in the release of viral particles from infected cells, making NA an attractive target of anti-influenza drugs (*37*). The amino acid sequences of NA differ by approximately 50% among the subtypes of influenza A virus. To understand the mechanisms of infection, it is critical to be able to predict the variable sites within the NA molecule. Based on the 3D structure of the NA, the predicted sequences correlated well with those obtained from isolated viral strains. The majority of the residues contacting the substrate, sialic acid, were consistent with those found in isolated viral strains.

We therefore discuss the relationship between amino acid properties and the characterization of the protein from the results of our analysis of the generated amino acid sequences.

## MATERIALS AND METHODS

To predict the amino acid sequences that were adaptable to the main-chain of a query protein, first of all, two Monte Carlo procedures on different layers were performed in this work (Fig. 1). A randomly generated amino acid type, first of all, was assigned to a query position (a residue site on the main-chain); the judgment if an amino acid was adaptable to the main-chain was made with a scoring function that was generated from a thermodynamic ensemble consisting of the possible side-chain conformations. To generate the ensemble, a side-chain library was constructed from a variety of side-chain conformations taken from known protein structures. The scoring function was defined as the free-energy difference ($\Delta\Delta G$) between the folded and unfolded states of the protein. Both the random

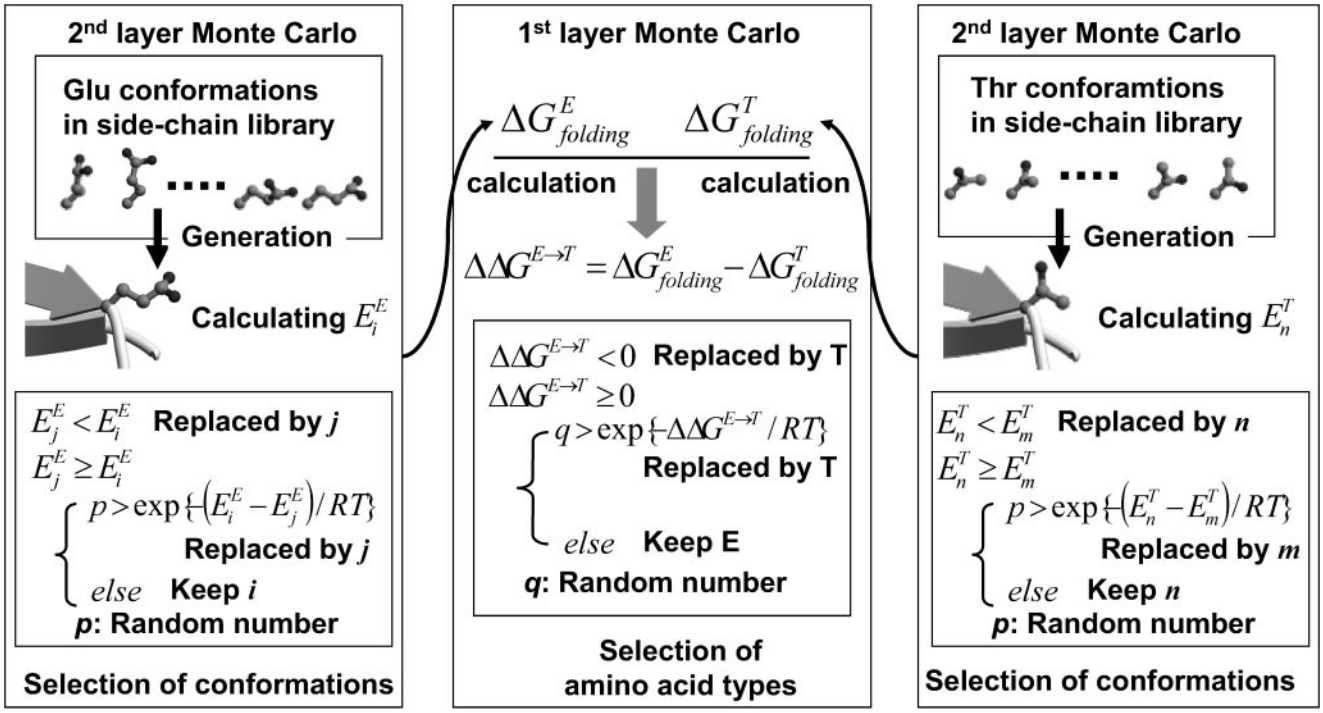


Fig. 1. **Two layers Monte Carlo procedures.** The 1st layer Monte Carlo procedure manages the selection of the side-chain conformation. In this procedure, the optimal side-chain conformation was selected. At the mean time, $\Delta G_{folding}^E$ value was calculated from the generated conformations. Analogously, $\Delta G_{folding}^T$ was calculated. The 2nd layer Monte Carlo procedure manages the selection of amino acid types using $\Delta\Delta G^{E\to T}$. If $\Delta\Delta G^{E\to T}$ was negative, $T$ was selected. If $\Delta\Delta G^{E\to T}$ was positive, a random number, $p$, was generated that was uniformly distributed between 0.0 and 1.0. If $p \geqslant \exp\{\Delta\Delta G^{E\to T}/RT\}$, $E$ was replaced by $T$. If not, $E$ was kept at that position.

selection of the amino acid type and the judgment of the side-chain adaptability to the main-chain were performed using the Metropolis Monte Carlo method. This procedure was executed repeatedly at every residue site. The details of the procedure follow.

*Procedure for Selection of Amino Acid Types*—Before beginning the sequence prediction, all residues on the given main-chain were set to Gly, and hydrogen atoms were added. First, the N-terminal residue served as the initial query position to which an appropriate amino acid should be assigned. An amino acid type at the query position was set by translating genetic code (codon)–based random sequences, which consisted of three-letter quaternary code (A, T, C, G) that avoided those sequences corresponding to stop codons. The codon was modulated to reproduce the frequency of amino acids occurring in wild-type proteins; in this sense, the codon was not entirely random.

Next we consider the change of amino acid type from α to β at the query position. At the initial prediction stage, α = Gly. To present the method in a general form, however, we assume that α can be any amino acid type. The free-energy difference with respect to folding for each amino acid is given as:

$$\Delta G^X_{folding} = G^X_{folded} - \Delta G^X_{reference}, \quad (1)$$

where the superscript $X$ designates α or β; when $X$ = α, the computation is performed for amino acid α, while when $X$ = β, it is computed for β. In this study, the unfolded state is referred to as the "reference state," from which the free-energy difference is measured. Then, $\Delta\Delta G^{\alpha\to\beta}$ is defined as

$$\Delta\Delta G^{\alpha\to\beta} = \Delta G^{\beta}_{folding} - \Delta G^{\alpha}_{folding} = \Delta G^{\beta\to\alpha}_{folded} - \Delta G^{\beta\to\alpha}_{reference}. \quad (2)$$

If $\Delta\Delta G^{\alpha\to\beta}$ is less than zero, amino acid β is thermodynamically more adaptable to the given main-chain than α. Note that $\Delta\Delta G^{\alpha\to\beta}$ is commonly used in free-energy perturbation studies, in which the computations are based upon a thermodynamic cycle. To select an appropriate amino acid at the query position, we used $\Delta\Delta G^{\alpha\to\beta}$ as the scoring function.

The free energy, $\Delta G^X_{folding}$ in Eq. 1 of the folded state was defined as:

$$G^X_{folded} = -kT\ln Z^X_{folded}, \quad (3)$$

where $k$ and $T$ are the Boltzmann constant and temperature, respectively. In this study, $T$ was set at 300 K in all the procedures. The $Z^X_{folded}$ is the partition function defined as:

$$Z^X_{folded} = \sum_{i=1}^{N} \exp\{-E^X_i/kT\}, \quad (4)$$

where $E^X_i$ is the potential energy of the system, $i$ specifies a randomly generated side-chain conformation for amino acid $X$, and $N$ is a number of the side-chain conformations generated. The mechanism by which the side-chain conformations are generated is described later. When α is Gly, the side-chain conformation is unique. The free energy, as currently defined, does not take into account the entropic effect from the main-chain conformational variety, because the main-chain conformation is fixed. $Z^X_{folded}$ is a quantity computed from the 2nd layer Monte Carlo sampling. As shown in "RESULT" section, the introduction of the 2nd layer Monte Carlo sampling improved the prediction accuracy. Note that if $N$ =1, $Z^X_{folded}$ is equal to $E^X_1$. This means that for $N$ = 1 the scoring function is reduced to a similar form with those used in other methods (*6*, *38*) because the scoring function is computed only from single side-chain conformation in this case.

$E^X_i$ was defined as:

$$E^X_i = E^X_{non\text{-}bonded} + E^X_{solv}, \quad (5)$$

where $E^X_{non\text{-}bonded}$ and $E^X_{solv}$ are non-bonded interactions and solvent energies, respectively, for the conformation $i$ of amino acid $X$. The parameters necessary to compute $E^X_{non\text{-}bonded}$ were taken from the AMBER force field (*39*) with a dielectric constant of $\varepsilon = 4r$ ($r$: atom–atom distance). ABMER is a molecular dynamic package and it is widely used in the field of computational biology. The electrostatic interaction was truncated at a cutoff distance of 12 Å, which is generally used in molecular dynamics and mechanics studies.

The solvent term $E^X_{solv}$ is important to obtain accurate free-energy differences between the folded and unfolded states. Several approximated solvent energies have been applied to the protein design issue (*40*, *41*). Here we used a distance-dependent dielectric model with surface area–dependent solvent energy, since this combination has often been used and is successful for non-polar residues in the protein design (*34*, *42*). Here, $E^X_{solv}$ was given by:

$$E^X_{solv} = \sum_{j=1}^{M} \sigma_j A_j \quad (6)$$

where $M$ is the number of atoms, and $A_j$ and $\sigma_j$ are the solvent-accessible surface area of the $j$th atom and its contribution to $E^X_{solv}$, respectively. The values for $\sigma_j$ were taken from the parameters set by Ooi and Oobatake (*43*).

The free energy $G^X_{reference}$ (Eq. 1) for the unfolded state was calculated in the similar way to $G^X_{folding}$. The exact description of the unfolded state, however, in which the polypeptide chain fluctuates in a wide conformational space, is difficult using this prediction method. Therefore, the unfolded state was approximated as a single residue at the query position, which consists of only the query position and neglects the other positions. The various side-chain conformations of amino acid $X$ were generated on the main-chain; the potential energy $E^X_i$ was calculated to obtain $G^X_{reference}$.

After calculating the scoring function $\Delta\Delta G$, the Metropolis Monte Carlo algorithm was used to determine if amino acid α could be replaced by β at the query position: if $\Delta\Delta G$ was negative, β was selected. If $\Delta\Delta G$ was positive, a random number, $p$, was generated that was uniformly distributed between 0.0 and 1.0. If $p \geqslant \exp\{-\Delta\Delta G/RT\}$, α was replaced by β. If not, α was retained at that position. Then, the temperature in this procedure was set at 300 K. When β was selected, the side-chain conformation of β was set as the lowest potential energy conformation of the generated solutions. This completed the procedure for the first query position.

Next, the query position shifted to the second residue; an adaptable amino acid was assigned to the position according to the procedure described above. To calculate $\Delta\Delta G$, the side-chain conformation of the first (*i.e.*, N-terminal) residue was fixed as the lowest potential energy solution determined above. After the same set of calculations outlined above, the side-chain conformation of the next query position was set to the lowest potential energy solution of the sampled conformations. The query position shifted to the third residue; an adaptable amino acid was assigned to this position, in which $\Delta\Delta G$ was computed after fixing the side-chain conformations of the first and second residues to those of the lowest potential energies determined before. This procedure was continued up to the C-terminal residue. The term "Monte-Carlo cycle" is used to specify such a set of trials spanning from the N- to the C-terminal residues.

After the first Monte-Carlo cycle, the sequence was altered to have side-chain conformations with the lowest potential energies. At this stage, the sequence had not yet converged to the optimal sequence for the scoring function because the side-chain conformations surrounding the query position is fixed. Then, the second Monte-Carlo cycle was started, and after the second Monte-Carlo cycle, the third Monte-Carlo cycle was executed, and so on. By performing many Monte-Carlo iterations, the sequences giving the side-chain conformations with the optimal energies will be found.

The generation of side-chain conformations (Eq. 4) used a side-chain library consisting of 9,350 side-chain conformations with tertiary coordinates for 18 amino acid types (except for Ala and Gly) extracted from known protein structures. This library consists of non-redundant side-chain conformations, but eliminated similar side-chain conformations having root mean square deviation (rmsd) values less than 0.2 Å different from each other.

*Application of the Method*—To assess the performance of this method, we first predicted the sequence of the core residues of two small proteins, SH3 domain (PDB entry = 1CKA) and Protein G (PDB entry = 1PGB) (*44*). These proteins have previously been used for protein design (*4*, *25*, *34*). The core residues, for which the solvent-accessible surface area was less than 30%, were chosen to be the same residues detailed by Wernisch *et al*. In our computations, all atoms, with the exception of those to be predicted (*i.e.*, the core residues), were fixed as the conformations present in the wild-type proteins throughout the design procedure. We performed 10 prediction procedures, each consisting of 1,000 Monte-Carlo cycles, yielding 10,000 sequences in total. For every sequence set, the first 100 amino acid sequences were discarded from the analysis, because these sequences were influenced by the initial trial sequence, in which all amino acids were Gly. The remaining 900 sequences were combined into one set of sequences, giving a total of 9,000 sequences. For these sequences, a non-redundant sequence set was created by removing equivalent sequences.

The method was then applied to the entire sequence of SH3, Protein G and bovine pancreatic trypsin inhibitor (BPTI; PDB entry = 5 PTI), with the exception of the Cys residues that formed disulfide bonds. Here, for each protein, we performed four prediction procedures, each consisting of 25,000 Monte-Carlo cycles, to yield 100,000 sequences in total. We also discarded the initial 100 sequences from each of the prediction procedures.

To assess the efficiency of the 2nd layer Monte Carlo procedure, two predictions were performed: one with and the other without the 2nd layer Monte Carlo procedure, in which the number of side-chain conformations, $N$, were set at 1 and 15, respectively. For the statistical assessment of the amino acid selectivity of our method, we calculated the distributions of sequence identities of the uniquely predicted sequences against the wild-type sequence. The sequence identity was computed excluding the Cys residues that formed disulfide bonds. We also calculated the sequence identities of the homologous sequences and a number of randomly generated sequences. We excluded proteins with sequence identities greater than 70% (to eliminate mutants) from the set of homologous proteins. The random sequences were generated from genetic code–based random numbers, as described above.

The second analysis is to measure the diversity of the amino acid type at each position using the probabilistic entropy, $S$, defined by

$$S^n(f) = -\sum_{k=1}^{20} f_k^n \ln f_k^n, \quad (7)$$

where $n$ and $k$ represent the position and the amino acid type, respectively. $f_k^n$ is the frequency ratio of amino acid $k$ at the position $n$ in the sequences. The range of $S$ is from 0 to ln20 (∼3.0). If a position has a large value of $S$, the variation of the amino acid types is large at the position.

Finally, we applied this method to the prediction of variable sites on the NA molecule of influenza virus. The prediction was performed for amino acid residues composing the ligand-binding sites of NA. The main-chain conformation of NA (PDB entry = 2QWB), which forms a complex with *O*-sialic acid (SIA), was used for the prediction. The force-field parameters (van der Waals and electrostatic parameters) for the SIA were obtained using the antechamber package (*45*). The atomic partial charges were computed with the am1bcc option (AM1 Mulliken charge). Residues with a distance shorter than 6.0 Å from the side-chain atoms to SIA within the complex were selected for the sequence prediction of NA. In this computation, we performed three prediction procedures, each consisting of 100,000 Monte-Carlo cycles to yield 300,000 sequences in total for the NA molecule. Again, we discarded the initial 100 sequences.

## RESULTS AND DISCUSSION

*Prediction of Core Residues*—We obtained 770 unique sequences for the 12 core residues (positions 4, 6, 10, 17, 18, 20, 26, 28, 39, 41, 49, and 54) of the SH3 domain from the 9,000 sequences generated by 10 prediction iterations. Eight of the twelve positions, with the exceptions of positions 6, 17 20, and 41, were dominated by the native residues in the predicted sequences (Table 1). The predicted models had similar hydrophobic interactions as those in the X-ray structure at these positions. We discuss the results that disagreed with the wild-type sequence below.

Positions 6 and 20 in the wild-type protein were Ala and Phe, respectively. The sixth position, however, was occupied by Met with the highest frequency (Table 1). Position

Table 1. **Amino acid frequency of predicted sequences in SH3 domain.**[a]

| Amino acid types | Position | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Val 4 | Ala 6 | Phe 10 | Asp 17 | Leu 18 | Phe 20 | Leu 26 | Ile 28 | Ala 39 | Asp 41 | Ile 49 | Val 54 |
| Ala | 0.22 | 0.33 | | **0.54** | | 0.34 | 0.09 | 0.06 | **0.77** | **0.37** | | 0.15 |
| Arg | | | | | | | | | | | | |
| Asn | | | | | | | | | | | | |
| Asp | 0.20 | | | 0.00 | | | | | | 0.00 | | 0.05 |
| Cys | 0.06 | | | | | | | | | | | |
| Gln | | | | | | | | | | | | |
| Glu | | | | | | | | | | | | |
| Gly | 0.06 | | | | | **0.43** | 0.10 | | 0.15 | 0.30 | 0.05 | |
| His | | | | | | | | | | | | |
| Ile | | | | | | | 0.17 | **0.60** | | | **0.49** | |
| Leu | | | | | **0.93** | | **0.23** | | | 0.09 | | 0.18 |
| Lys | | | | 0.42 | | | | | 0.05 | 0.10 | | |
| Met | | **0.39** | | | | | 0.15 | | | | | |
| Phe | 0.10 | 0.25 | **0.93** | | | 0.22 | | | | | | 0.10 |
| Pro | | | | | | | | | | | | |
| Ser | | | | | | | | | | | | |
| Thr | | | | | | | | | | | | |
| Trp | | | | | | | | | | | | |
| Tyr | | | | | | | | | | | | |
| Val | **0.31** | | | | | | 0.19 | 0.30 | | 0.07 | 0.43 | **0.44** |

[a]Frequency of amino acid types are shown. The values in bold type represent the highest frequency. The underlined values are the wild type amino acids.

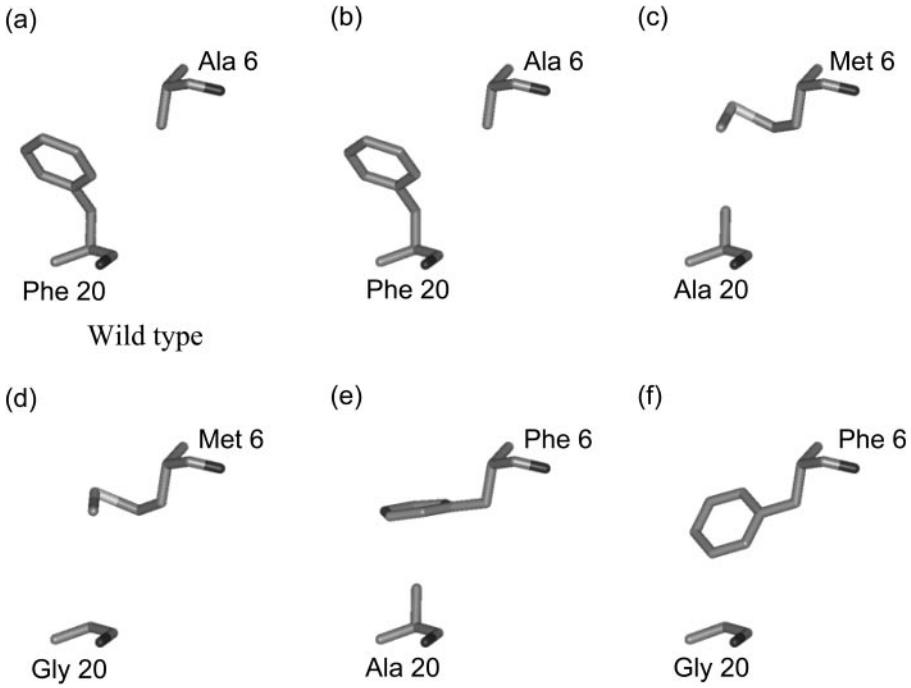


Fig. 2. **Comparison of the side-chain conformations in the X-ray structure with those from predicted positions 6 and 20 of the SH3 domain.** (a) Conformations of Ala 6 and Phe 20 in the X-ray structure. (b)–(f) Predicted conformations at position 6 and 20 are shown from the same viewpoint as that in (a).

20 was frequently occupied by Gly. These two amino acids create a hydrophobic pair in the predicted structure, which provides the foundation for the frequent generation of the following five pairs of amino acids: (Ala, Phe), (Met, Ala), (Met, Gly), (Phe, Ala) and (Phe, Gly), in which the former is the amino acid at position 6, and the later is at position 20. Note that the first pair is that found in the wild-type protein. Considering the side-chain conformations, all of these pairs exhibited a well-packed conformation within the surrounding atoms in the X-ray structure (Fig. 2). In the two pairs of (Ala, Phe) and (Phe, Ala), the benzene ring of Phe always adopted the same position

Table 2. **Amino acid frequency of predicted sequences in Protein G.**[a]

| Amino acid types | Position | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Tyr 3 | Leu 5 | Leu 7 | Ala 20 | Ala 26 | Phe 30 | Ala 34 | Val 39 | Phe 52 | Val 54 |
| Ala | | 0.05 | | **0.78** | **1.00** | | **0.85** | | | |
| Arg | | | | | | | | | | |
| Asn | | | | | | | | | | |
| Asp | | | | | | | | | | |
| Cys | | | | | | | 0.07 | | | |
| Gln | | | | | | | | | | |
| Glu | | | | | | | | | | |
| Gly | | | | 0.15 | | | 0.08 | | | |
| His | | | | | | 0.12 | | | | |
| Ile | | | **0.44** | | | | | **0.46** | | |
| Leu | | **0.95** | 0.22 | | | 0.12 | | 0.14 | | |
| Lys | | | | | | | | | | |
| Met | | | 0.14 | | | 0.05 | | 0.12 | | |
| Phe | **0.93** | | | | | **0.71** | | | **1.00** | |
| Pro | | | | | | | | | | |
| Ser | | | | | | | | | | |
| Thr | | | | | | | | | | |
| Trp | | | | | | | | | | |
| Tyr | 0.07 | | | | | | | | | |
| Val | | | 0.20 | | | | | 0.29 | | **0.95** |

[a]See legend in Table 1.

by sharing the same space (Fig. 2a and b). Homologous sequences, identified by a BLAST search of a non-redundant sequence database (NRDB: available from ftp://ftp.embl-heidelberg.de/pub/databases/nrdb/nrdb), did not contain the (Phe, Ala) pair in the SH3 domain family. Instead, the pair from the wild type is well conserved throughout the family. We presume that this mutation is structurally possible, but evolutionarily unselected, in the family. Such a mutation, however, could be exploited in artificial protein design.

Position 17 was occupied by Ala or Lys, while position 41 was filled by Ala or Gly, despite the presence of Asp at both positions in the wild-type protein. In the X-ray structure, OD1 of Asp41 forms two H-bonds with the N of Glu 43 (3.01 Å) and the N of Lys 45 (3.16 Å); Asp41 is buried inside the protein. Our scoring function tends to be misleading for buried hydrophilic residues; the value of the solvent energy for the reference state, which was computed from a single-residue representation of the query position, is markedly lower than that for the folded state. In addition, van der Waals interactions with the environmental atoms under tight packing yield a low potential value. As a result, small residues such as Ala or Gly, and occasionally Val and Leu, are preferably assigned to the buried residues, instead of the amino acids seen in the wild-type sequences.

By the prediction of the core residues of Protein G, 59 unique sequences were identified from the 9,000 sequences generated by this method. The number of unique sequences was considerably smaller than that seen for SH3, indicating that the sequences converged to a narrow solution set. The majority of the core positions were occupied by the same amino acid residues as those seen in the wild-type protein (Table 2). The frequencies of amino acid correlation at positions 5, 26, 52 and 54 were greater than 90%. These results indicate that the amino acid sequences converged both evolutionarily and physicochemically into that seen in the wild-type protein.

Only poor agreement was observed at position 3, which is a Tyr residue in the wild-type sequence (Table 2). At position 3, Phe was predicted with a high frequency (97%), instead of Tyr. In the X-ray structure, Tyr 3 is buried within a hydrophobic core consisting of Ala 20, Ala 26, and Phe 30. The OH group of Tyr 3 does not form a hydrogen bond with the surrounding atoms of the protein interior. The ratio of solvent-accessible surface area in the buried state to that in the solvent-exposed extended conformation was 3% using a 1.4 Å water probe (*46*). As the predicted sequences at positions 20, 26, and 30 coincided with those seen in the wild-type protein with high frequencies, the environment surrounding Tyr 3 appears to be similar to that of the wild-type protein. At position 3, Phe is likely to be more adaptable than Tyr for the given main-chain conformation.

It is difficult to account for evolutionary issues in this prediction method. The results, however, demonstrate that this method is useful for the prediction or modeling of structurally and energetically adaptable amino acid sequences.

*Prediction of the Entire Sequence*—The sequence variety from the prediction for the entire protein is considerably different from that of the core residue regions (Table 3). Over 70,000 unique sequences were obtained from the 99,600 sequences predicted by the methods with $N = 1$, whereas for $N = 15$, over 90,000 sequences were obtained in a similar fashion (Table 3). These unique sequences exhibited a high similarity with each wild-type protein, and are close to 25% identity. Especially, the average sequence identity for SH3 domain (1CKA) was nearly 30%. These values ensure that our method sufficiently worked to determinate proteins with the same fold (*47*).

Table 3. **Results of the predicted sequences in the whole prediction.**[a]

| Name of Proteins | PDB entry | $N = 1$ | | | $N = 15$ | | | Random | |
|---|---|---|---|---|---|---|---|---|---|
| | | Number of sequences | Sequence identity (%) | | Number of sequences | Sequence identity (%) | | Sequence identity (%) | |
| | | | maximum | average | | maximum | average | maximum | average |
| SH3 domain | 1CKA | 77,328 | 49.1 | 27.8 | 97,223 | 50.9 | 30.3 | 15.8 | 5.0 |
| Protein G | 1PGB | 70,705 | 42.9 | 24.3 | 97,903 | 39.3 | 21.4 | 16.1 | 5.1 |
| BPTI | 5PTI | 77,965 | 34.6 | 21.7 | 98,613 | 42.3 | 25.6 | 17.3 | 5.7 |

[a]The sequence were predicted by two methods; one without and other with the 2nd layer Monte Carlo procedure, in which the number of side-chain conformations, $N$, were set at 1 and 15, respectively.

For individual proteins, the maximum sequence identities for all three proteins are over 30% and 1CKA is the largest at 50.9%. In general, the proteins with a sequence identity greater than 40% not only share the same folding pattern but also have similar tertiary structure (*18*). Therefore the predicted sequences possibly exhibit the same fold as the wild-type proteins.

Figure 3 shows the distributions of the sequence identity for the predicted sequences against the wild-type protein. The distribution of the homologous proteins for 1PGB does not appear in Fig. 3 due to a lack of homologous proteins in the sequence database. The distribution of the predicted sequences by the method with $N = 15$ for 1CKA and 5PTI were significantly overlapped onto those of the homologous proteins. The distributions for $N = 1$ showed similar overlapping regions, but somewhat smaller. The sequences distributing in the overlapping regions would likely have the same fold as the wild-type protein with a greater probability than those distributing outside the overlapping regions. Therefore, majority of the predicted sequences from our method would be shown to exhibit the same fold as the wild-type proteins. On the other hand, the distributions of the random sequences were clearly separate from those of the predicted and homologous sequences. We conclude that the sequences did not converge on the wild-type sequence by random changes and the main-chain constraint guides the sequences toward the wild-type sequence.

*Efficiency of the 2nd Layer Monte Carlo Procedure*—A large part of the overlapped region is observed between the two distributions of the methods with $N = 1$ and $N = 15$ (Fig. 3). The average sequence identities predicted by the method with $N = 15$ were apparently better than that with $N = 1$ for 1CKA (Fig. 3a) and 5PTI (Fig. 3c), although sequence identities for 1PGB were similar between the two methods. Furthermore, when the amino-acid sequence was reduced to four amino-acid groups by their physicochemical properties (see footnote for Table 4), the reduced sequence identities from $N = 15$ were larger than those from $N = 1$ (Table 4). For core residues, the 2nd layer Monte Carlo procedure improved the reduced sequence identity for each protein. These results suggest that the 2nd layer Monte Carlo procedure effectively performed on our test proteins. The average values of the sequence identities on the method with $N = 1$ were similar to those reported in previous studies (*33*). Therefore, we expect that our method will produce amino acid sequences more similar with those of wild-type proteins than the previous methods do; especially on the reduced sequence identities calculated for the four physicochemical amino-acid groups.

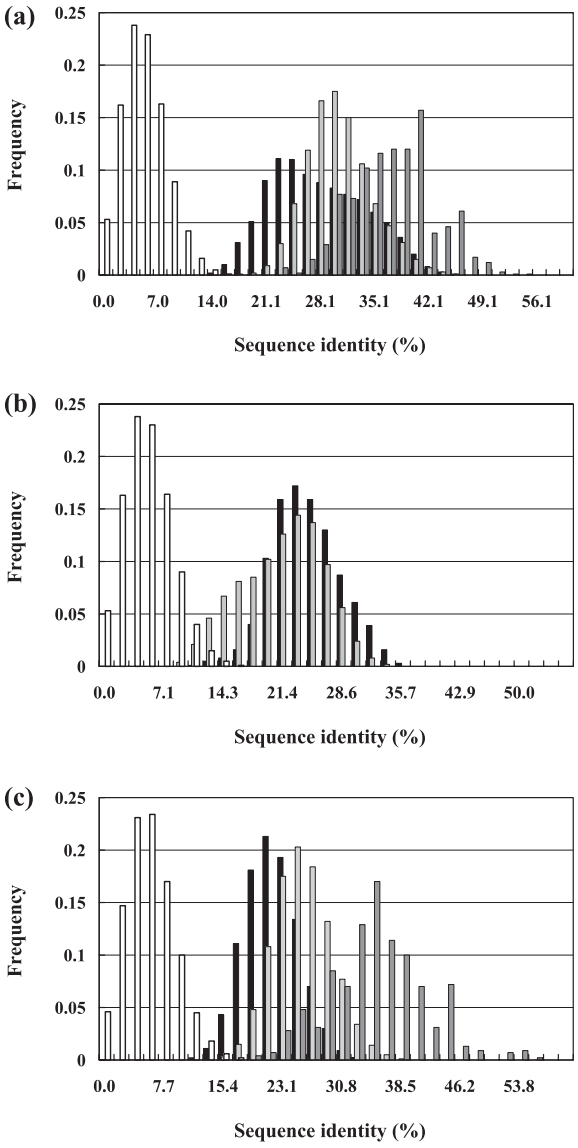


Fig. 3. **Distributions of the sequence identity of predicted sequences of (a) 1CKA, (b) 1PGB and (c) 5PTI for the wild-type sequence.** The black and light gray bars represent the predicted protein sequences with $N = 1$ and $N = 15$, respectively. The dark gray and white bars indicate homologous and random protein sequences, respectively. The histogram of the homologous proteins in the 1PGB was hidden in (b) because the number of homologous protein detected from the sequence databases was not enough to compare the other distributions. The sequence identities were calculated for all the positions, with the exception of those Cys residues forming disulfide bonds.

Table 4. **Sequence identity of the predicted sequences (%).**[a]

| Name of proteins | PDB entry | $N = 1$ | | | | $N = 15$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | All | | Core[b] | | All | | Core[b] | |
| | | maximum | average | maximum | average | maximum | average | maximum | average |
| SH3 domain | 1CKA | 49.1 | 27.8 | 72.2 | 34.2 | 50.9 | 30.3 | 77.8 | 42.0 |
| | | 66.7 | 48.8 | 94.4 | 70.2 | 66.7 | 49.4 | 94.4 | 77.5 |
| Protein G | 1PGB | 42.9 | 24.3 | 73.3 | 49.4 | 39.3 | 21.4 | 80.0 | 45.7 |
| | | 58.9 | 41.3 | 86.7 | 72.5 | 57.1 | 41.6 | 86.7 | 73.3 |
| BPTI | 5PTI | 34.6 | 21.7 | 42.9 | 23.6 | 42.3 | 25.6 | 57.1 | 30.8 |
| | | 59.6 | 38.8 | 64.3 | 42.0 | 63.5 | 43.8 | 71.4 | 45.4 |

[a]The sequence were predicted by two methods; one without and other with the 2nd layer Monte Carlo procedure, in which the number of side-chain conformations, $N$, were set at 1 and 15, respectively. The results show the sequence identity calculated for 20 amino acid (upper part) and for 4 groups on the basis of their physicochemical properties: an acidic group (Asp and Glu), a polar group (Asn, Gln, Ser, Thr, Tyr, and Cys), a basic group (Lys, Arg, and His), and a non-polar group (Trp, Phe, Gly, Ala, Val, Leu, Ile, Pro, and Met) (lower part).
[b]Core residues were defined those of which the solvent accessible surface area is less than 25%.

The probabilistic entropy values, $S$, in the prediction with $N = 1$ tends to have the larger values than those with $N = 15$ for all the query sites (Fig. 4). The predicted sequences with $N = 1$ showed the most diversity of amino acid type. The values of $S$ with $N = 1$ in the vicinity of the C-terminal region of the 1CKA and 1PGB have large values. In particular, the amino acid types at residues 50 and 52 in 1CKA and the residues 45 and 47 in 1PGB with $N = 1$ vary drastically due to the external physical location. However, these position show the small $S$ values with $N = 15$. Regarding the residues located on the surface of the proteins, a number of side-chain conformations were accepted from the 2nd layer Monte Carlo procedure sampling. On the other hand, the optimal side-chain conformation could hardly be detected by the predictions with $N = 1$. Therefore, highly flexible surface side-chain conformations having high scoring values give likelihood to change the amino acid types having low scoring values.

These results show that the 2nd layer Monte Carlo procedure could select the optimal side-chain conformation from N ones and provided the stable conformation. Less flexible side-chain conformations prevent the likelihood of changing the side-chain types. Consequentially, the predicted sequences by the method with $N = 15$ were similar to the wild-type ones.

*Prediction of Variable Sites in NA Molecule of Influenza Virus*—To understand the mechanisms of viral infection, it is important to predict the variable sites in the NA molecule. We applied our method to the *O*-sialic acid (SIA)–binding region of the NA molecule, for which the main-chain conformation of the R292K mutant (PDB: 2QWB) was used.

We obtained 1,467 unique sequences from the 300,000 sequences generated. The amino acid diversity of the predicted sequences tended to be larger than that seen in the sequences from the isolated wild-type viral strains (Table 5). The majority of positions had more than four candidate amino acids, including the wild-type amino acid. Ten of the 11 amino acid residues that were conserved among viral subtypes were consistent with the predicted ones (Table 5). Three Arg residues at positions 118, 292, and 371 form a pocket (*26*). Figure 5 illustrates this core interaction site between SIA and NA (2QWB where Arg292 was mutated to Lys). Position 118 was predicted as Arg or Lys, position 292 as Arg, Glu, and Lys, and position 371 was determined to be Arg. Arg was always included as one

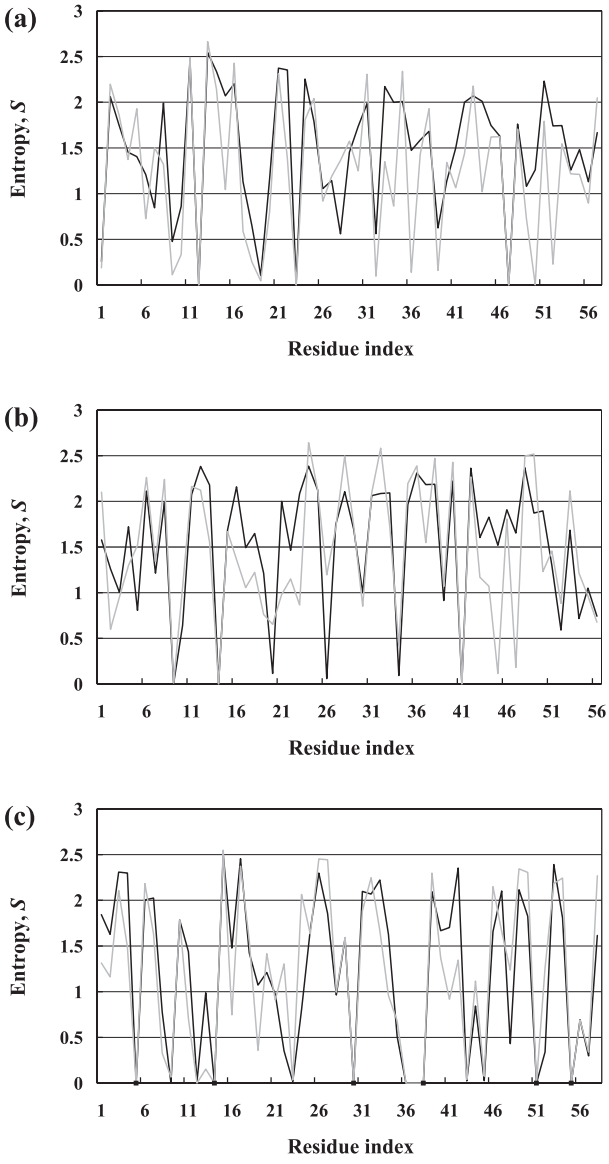


Fig. 4. **Probabilistic entropy for each position of (a) 1CKA, (b) 1PGB and (c) 5PTI.** And the line graphs colored in black and gray were the values of $S$ on the predictions with $N = 1$ and $N = 15$, respectively. The values of $S$ at the Cys residues forming disulfide bonds in 5PTI were set at 0.

Table 5. **Comparison of the amino acid types in the wild type proteins with the predicted sequences.**[a]

| Neuraminidase molecule | | | | | |
|---|---|---|---|---|---|
| Number[b] | Wild type | Prediction[c] | Number[b] | Wild type | Prediction[c] |
| **118** | R | **R**K | 245 | AS | R |
| **119** | E | ND**E**S | **274** | H | N**H**K |
| **134** | L | RQEH | **276** | E | DS |
| **151** | D | **D**E | **277** | E | RK |
| **152** | R | **R** | **292** | R | **R**EK |
| 156 | R | **R**NT | 294 | N | **N**DH |
| **178** | W | DEHY | 347 | NHPY | R |
| 179 | S | D**S**T | 348 | G | **G** |
| 222 | I | RDE | 371 | R | **R** |
| 223 | R | **R** | 406 | Y | D |
| **227** | E | ND**E**G | 427 | I | R |

[a]Amino acid types having a frequency over 5%. The sequences were extracted from a non-redundant sequence database (NRDB).
[b]PDB-based number. Conserved amino acid residues among subtypes are shown by underlined letters in bold.
[c]Amino acid types in the predicted sequences.

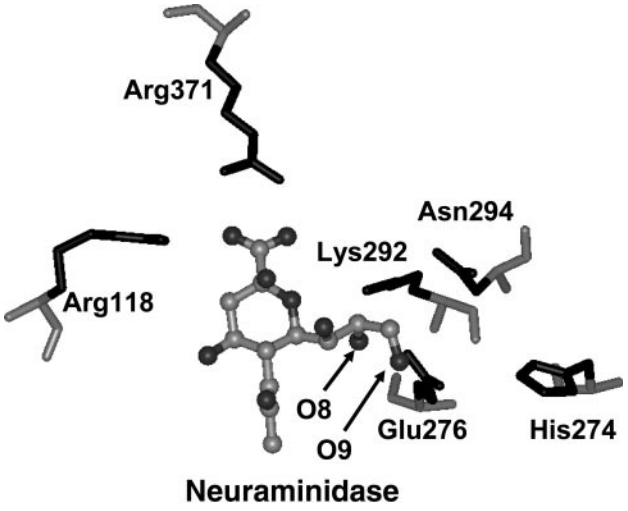


Fig. 5. **NA-SIA complex.** The SIA molecule is shown with the environmental residues of NA (PDB-entry: 2QWB).

of the predicted amino acids at each of the three positions. In the predicted sequences, position 292 had a different charged residue (Glu) from that seen in the wild type (Arg). In the predicted structure, this position may interact with the O8 atom in the SIA molecule, possibly via the COOH proton of Glu. These differences in amino acid residues from the wild-type residues in viral proteins may be interpreted as mutations that disappeared in the evolutionary process or as sequences that may emerge as structurally and physicochemically adaptable mutations in the future.

In conclusion, we have developed a method to predict the amino acid sequences that are adaptable to a given main-chain conformation. The predicted sequences correlated well with the wild-type sequences. The structures of the predicted amino acid sequences also exhibited coordinated mutations between large and small residues. The side-chain conformations of these sequences demonstrated close packing with the main-chain atoms, indicating that the structures of these predicted sequences may form the same folding conformations as the wild-type proteins. In our prediction of the NA molecule in a complex with an SIA molecule, we could obtain similar amino acid types as those seen in the wild type, especially those residues contacting the SIA molecule. The predicted amino acid types at a number of query positions, however, were not identified in the homologous sequences identified from a homology search using BLAST. We hypothesize that, while these mutant sequences were structurally selected in our computations, they were not functionally selected in nature. Therefore, there is a possibility that these sequences might appear in the future. The structure of the predicted sequences should be stable, because the scoring functions corresponding to $\Delta\Delta G$ were computed to be favorable; only those sequences adaptable to the backbone structure were obtained. This method could be useful to identify important residues even in a highly variable viral protein. This method may allow us to develop novel drugs and vaccines by predicting the variations within a target molecule.

## REFERENCES

1. Duan, Y. and Kollman, P.A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744
2. Snow, C.D., Nguyen, H., Pande, V.S., and Gruebele, M. (2002) Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106
3. Bryson, J.W., Desjarlais, J.R., Handel, T.M., and DeGrado, W.F. (1998) From coiled coils to small globular proteins: design of a native-like three-helix bundle. *Protein Sci.* **7**, 1404–1414
4. Dahiyat, B.I., Gordon, D.B., and Mayo, S.L. (1997) Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337
5. Dahiyat, B.I. and Mayo, S.L. (1996) Protein design automation. *Protein Sci.* **5**, 895–903
6. Dahiyat, B.I. and Mayo, S.L. (1997) De novo protein design: fully automated sequence selection. *Science* **278**, 82–87
7. Dahiyat, B.I. and Mayo, S.L. (1997) Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **94**, 10172–10177
8. Dahiyat, B.I., Sarisky, C.A., and Mayo, S.L. (1997) De novo protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789–796
9. Desjarlais, J.R. and Clarke, N.D. (1998) Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.* **8**, 471–475

10. Desjarlais, J.R. and Handel, T.M. (1995) De novo design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006–2018
11. Desjarlais, J.R. and Handel, T.M. (1995) New strategies in protein design. *Curr. Opin. Biotechnol.* **6**, 460–466
12. Desjarlais, J.R. and Handel, T.M. (1999) Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318
13. Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., and Kim, P.S. (1998) High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467
14. Hayes, R.J., Bentzien, J., Ary, M.L., Hwang, M.Y., Jacinto, J.M., Vielmetter, J., Kundu, A., and Dahiyat, B.I. (2002) Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. USA* **99**, 15926–15931
15. Kortemme, T., Ramirez-Alvarado, M., and Serrano, L. (1998) Design of a 20-amino acid, three-stranded beta-sheet protein. *Science* **281**, 523–526
16. Kuhlman, B. and Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA* **97**, 10383–10388
17. Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368
18. Kuhlman, B., O'Neill, J.W., Kim, D.E., Zhang, K.Y.J., and Baker, D. (2001) Conversion of monomeric protein L an obligate dimer by computational protein design. *Proc. Natl. Acad. Sci. USA* **98**, 10678–10691
19. Malakauskas, S.M. and Mayo, S.L. (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475
20. Ogata, K., Jaramillo, A., Cohen, W., Briand, J.P., Connan, F., Choppin, J., Muller, S., and Wodak, S.J. (2003) Automatic sequence design of major histocompatibility complex class I binding peptides impairing CD8+ T cell recognition. *J. Biol. Chem.* **278**, 1281–1290
21. Shimaoka, M., Shifman, J.M., Jing, H., Takagi, J., Mayo, S.L., and Springer, T.A. (2000) Computational design of an integrin I domain stabilized in the open high affinity conformation [see comments]. *Nat. Struct. Biol.* **7**, 674–678
22. Street, A.G., Datta, D., Gordon, D.B., and Mayo, S.L. (2000) Designing protein beta-sheet surfaces by Z-score optimization. *Phys. Rev. Lett.* **84**, 5010–5013
23. Street, A.G. and Mayo, S.L. (1999) Computational protein design. *Struct. Fold. Des.* **7**, R105–109
24. Strop, P., Marinescu, A.M., and Mayo, S.L. (2000) Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Sci.* **9**, 1391–1394
25. Su, A. and Mayo, S.L. (1997) Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6**, 1701–1707
26. Colman, P.M., Hoyne, P.A., and Lawrence, M.C. (1993) Sequence and structure alignment of paramyxovirus hemagglutinin-neuraminidase with influenza virus neuraminidase. *J. Virol.* **67**, 2972–2980
27. Voigt, C.A., Gordon, D.B., and Mayo, S.L. (2000) Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789–803
28. Cootes, A.P., Curmi, P.M.G., and Torda, A.E. (2000) Automated protein design and sequence optimisation: Scoring functions and the search problem. *Curr. Protein Pept. Sci.* **1**, 255–271
29. Mendes, J., Guerois, R. and Serrano, L. (2002) Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **12**, 441–446
30. Eisenmenger, F., Argos, P., and Abagyan, R. (1993) A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* **231**, 849–860
31. Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I. (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542
32. Koehl, P. and Levitt, M. (1999) De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161–1181
33. Raha, K., Wollacott, A.M., Italia, M.J., and Desjarlais, J.R. (2000) Prediction of amino acid sequence from structure [in process citation]. *Protein Sci.* **9**, 1106–1119
34. Wernisch, L., Hery, S., and Wodak, S.J. (2000) Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.* **301**, 713–736
35. Rao, S.N., Singh, U.C., Bash, P.A., and Kollman, P.A. (1987) Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature* **328**, 551–554
36. Kono, H. and Saven, J.G. (2001) Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J. Mol. Biol.* **306**, 607–628
37. Garman, E. and Laver, G. (2004) Controlling influenza by inhibiting the virus's neuraminidase. *Curr. Drug Targets* **5**, 119–136
38. Gordon, D.B., Marshall, S.A., and Mayo, S.L. (1999) Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513
39. Weiner, S.J., Kollman, P.A., Nguyen, D.T., and Case, D.A. (1986) An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **7**, 230–252
40. Pokala, N. and Handel, T.M. (2004) Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **13**, 925–936
41. Vizcarra, C.L. and Mayo, S.L. (2005) Electrostatics in computational protein design. *Curr. Opin. Chem. Biol.* **9**, 622–626
42. Street, A.G. and Mayo, S.L. (1998) Pairwise calculation of protein solvent-accessible surface areas. *Folding Design* **3**, 253–238
43. Ooi, T., Oobatake, M., Nemethy, G., and Scheraga, H.A. (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA* **84**, 3086–3090
44. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242
45. Wang, J., Wang, W., Kollman, P.A., and Case, D.A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, in press
46. Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400
47. Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. (1992) A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1698