PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

An efficient conformational sampling method for homology modeling

Rongsheng Han,[1]* Alejandra Leo-Macias,[1†] Daniel Zerbino,[1‡] Ugo Bastolla,[1] Bruno Contreras-Moreira,[2] and Angel R. Ortiz[1§]

[1] Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, Madrid, Spain

[2] Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

ABSTRACT

The structural refinement of protein models is a challenging problem in protein structure prediction (Moult et al., Proteins 2003;53(Suppl 6):334–339). Most attempts to refine comparative models lead to degradation rather than improvement in model quality, so most current comparative modeling procedures omit the refinement step. However, it has been shown that even in the absence of alignment errors and using optimal templates, methods based on a single template have intrinsic limitations, and that refinement is needed to improve model accuracy. It is thought that failure of current methods originates on one hand from the inaccuracy of the effective free energy functions adopted, which do not represent properly the energetic balance in the native state, and on the other hand from the difficulty to sample the high dimensional and rugged free energy landscape of protein folding, in the search for the global minimum. Here, we address this second issue. We define the evolutionary and vibrational armonics subspace (EVA), a reduced sampling subspace that consists of a combination of evolutionarily favored directions, defined by the principal components of the structural variation within a homologous family, plus topologically favored directions, derived from the low frequency normal modes of the vibrational dynamics, up to 50 dimensions. This subspace is accurate enough so that the cores of most proteins can be represented within 1 Å accuracy, and reduced enough so that Replica Exchange Monte Carlo (Hukushima and Nemoto, J Phys Soc Jpn 1996;65:1604–1608; Hukushima et al., Int J Mod Phys C: Phys Comput 1996;7:337–344; Mitsutake et al., J Chem Phys 2003;118:6664–6675; Mitsutake et al., J Chem Phys 2003;118:6676–6688) (REMC) can be applied. REMC is one of the best sampling methods currently available, but its applicability is restricted to spaces of small dimensionality. We show that the combination of the EVA subspace and REMC can essentially solve the optimization problem for backbone atoms in the reduced sampling subspace, even for rather rugged free energy landscapes. Applications and limitations of this methodology are finally discussed.

Proteins 2008; 71:175–188.

Key words: comparative modeling; conformational search; multiple structure analysis; principal components analysis; normal mode analysis; Replica Exchange Monte Carlo.

INTRODUCTION

With the steady progression of structural genomics projects,[1] comparative modeling[2–4] is becoming an increasingly important technique for building protein structural models[5] and their complexes.[6,7] The latest CASP editions[8–12] have witnessed continuous progress in the quality of the sequence to structure alignments, a key step in the process of building homology models.[13] Oddly, the analysis of the same CASP results has also

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.
Grant sponsor: Education and Science Ministry of Spain; Grant numbers: BIO2001-3745, BIO2005-0576, GEN2003-206420-C09-08; Grant sponsor: Comunidad de Madrid; Grant numbers: GR/SAL/0306/2004, 200520M157; Grant sponsor: CSIC intramural program; Grant number: PIF2005; project CAR; Grant sponsor: BBVA Foundation; Grant sponsor: Ramón Areces Foundation.
*Current address: Department of Mathematics and Physics, North China Electric Power University, Beijing, 102206, People's Republic of China.
†Current address: European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, Heidelberg, Germany.
‡Current address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 ISD, United Kingdom.
§Correspondence to: Angel R. Ortiz, Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, Madrid, Spain. E-mail: aro@cbm.uam.es
Received 6 February 2007; Revised 23 May 2007; Accepted 8 June 2007
Published online 11 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21672

indicated that this progress was not directly transformed into an obvious improvement in the quality of the final three-dimensional structures.[9] This suggests that substantial work still has to be devoted to the development of methods aimed at the refinement of these models.[4,6] In fact, as it has been pointed out[14,15] if this tendency is to continue, refinement upon identification of a sequence-structure alignment may soon become a major bottleneck for turning predicted models useful for biology. Refinement becomes particularly relevant when the sequence identity between target and template falls below 30%. Current models in this range generally have root mean square deviations (RMSD) above 2.0 Å. This is also the most frequent case in a modeling project; it has been estimated that the probability that the query sequence shares <30% of identity to a known structure of the same fold is at least 50%.[3] Improving the accuracy of comparative models in this regime is truly important, because applications such as virtual drug discovery,[16,17] molecular replacement,[18–20] or the extraction of functional properties,[21,22] to name just a few applications, all crucially depend on deriving accurate models. However, and in spite of some anecdotal evidence using molecular dynamics or Monte Carlo simulations,[23,24] no reliable method has emerged yet to deal with this situation.

Difficulties with refinement are thought to originate from the existing force fields, not accurate enough to reproduce the balance of forces in the native state, and from the need to approach the global minimum of a highly dimensional and rugged energy landscape.[25] But a deeper reason may also be brought into consideration. It is important to realize that in refining comparative models the goal is to simulate an evolutionary process, one that takes us from the template to the target, not a physical one. However, typical refinement methods, such as those based on straightforward use of molecular dynamics simulations with typical force fields are meant to simulate physical processes and, therefore, are not designed to handle evolutionary transformations. Although Leo-Macias *et al.*[26] discovered recently that the evolutionary variation of protein cores in protein families proceeds along a combination of a small number of low-frequency modes imposed by the topology, it is easy to imagine situations where straightforward evolutionary paths between template and target may need to transverse high energy regions in the target phase space.

A more successful approximation to this problem might be to better understand the nature of evolutionary transformation and to implement the insights gained into new modeling methods. This is the concept, recently explored by Qian *et al.*,[27] with encouraging results. They applied energy-based refinement in reduced sampling subspaces, adopting as degrees of freedom of the protein backbone the main principal components (PC) of the observed structural variation in structural alignments[28] of proteins homologous to the target. Both Qian *et al.*[27] and Leo-Macias *et al.*[26] had showed, studying families of homologues structures, that the most relevant part of the evolutionary subspace is low dimensional. PC sampling therefore could afford a huge reduction of degrees of freedom with minimal loss in the maximum accuracy attainable, because the PCs represent the most likely evolutionary movements of the protein chain. Moreover, the outcome should be less sensitive to the inaccuracy of the energy function, because the reduction of the sampling space eliminates false attractors. Low-energy models in this subspace were identified with the Rosetta high-resolution energy function, and they consistently had lower RMSD to the native backbone than the starting templates. Although these results were encouraging, the overall improvement was modest, of the order of 0.3 Å on the average. It was not clear to what extent this limitation was due to the inaccuracy of the force field, the fact that Qian *et al.* employed only three PCs in their computations, or due to the sampling limitations of the Monte Carlo algorithm employed. On the other hand, and owing to the high variance in prediction quality observed in the series, the concern arises as to whether the number and diversity of structures available may seriously affect the accuracy and robustness of the method. If the pool of homologous structures is small, the evolutionary sampling represented in the multiple structure alignment might be poor, and in these conditions it is unlikely that the target protein can be accurately modeled in the PC subspace. We recently showed[26] that low-frequency normal modes (NMs) carry relevant information that can complement the definition of the sampling subspace in those cases where the structural sampling available for the PC analysis is limited, and can therefore help to stabilize the variance observed in model quality and reduce the dependency of the number of available homologous structures.

Here, we address those open questions. Our first task was to disentangle the influence of the sampling subspace from that of the energy function and the search algorithm. We first evaluated the quality of sampling subspaces spanned by different numbers of PCs and NMs by analytically projecting the protein structures onto different subspaces and measuring the RMSD between the projected structure and the native one. We found that the subspace with all the evolutionary PCs and NMs up to about 50 dimensions appears to satisfy the two conflicting requirements of reasonable geometric accuracy and limited volume, so that sampling algorithms can be able to explore it efficiently. Then we turned our attention to the sampling algorithm and studied whether efficient search engines, particularly Replica Exchange Monte Carlo (REMC), can be successfully applied. REMC[29] is one of the best current methods for optimization in rugged landscapes, but its applicability to proteins has been limited by the rapid increase of computational requirements for increasing number of degrees of freedom. We considered two energy functions with well defined, known global minima. First, we considered as energy function the root mean square deviation (RMSD) between the sampled and native

structure. However, this is an exceedingly simple function, with a smooth landscape. We then devised an energy function with mouldable ruggedness and analytically known global minimum. We will show that the method consistently reaches the global minimum basin, even for realistic conditions of ruggedness.

METHODS

Protein sets

Three data sets were used in this work. The first one consists of 547 proteins retrieved from the ASTRAL40 database,[30] corresponding to 30 large, well-studied superfamilies covering the most important classes of the SCOP database[31] (all α proteins, all β proteins, α/β proteins, α + β proteins). The maximum percentage of identity between two proteins within each superfamily is 40%. See Table I of the supplementary material for details.

The second and third sets were used to compare our results with those obtained through state of the art homology modeling protocols. The second set is a collection of 67 CASP5 targets and the third one is the set of the best models that participant groups generated for those targets. To model the targets, we built a library of PDB templates which includes 6182 PDB chains from the PDB-SELECT[32] 90% nonredundant set of PDB chains corresponding to April 2002. For each CASP5 target, we then applied the following steps: (1) retrieve with MAMMOTH[33] a list of structurally similar proteins from the PDB library, defined as those structures with a MAMMOTH score above 4. (2) Discard templates too long or too short (lengths of 2.5 or 0.4 times the length of the target). (3) Select the optimum set of templates. To this end, we applied a Metropolis Monte Carlo optimization to the set of retrieved structures to maximize the size of the core of the structural alignment (see below). This optimization step was needed because in some cases the core of the structural alignment was vanishingly small. For all the targets, at least three templates were found and more than 60% of the total structure was modeled on average. Only in two cases, T0141 and T0184_1, this percentage was lower than 30%, because of the difficulty for finding templates for these targets. See Table II of the supplementary information for more details.

Evolutionary subspace

Principal component analysis (PCA)

Each family of structures was aligned with the multiple alignment program MAMMOTH-mult.[28] The strict core of the protein family is defined as the set of positions for which the Cα atoms are present in all proteins, and their pairwise distances are always smaller than 4 Å. For all positions in the strict core, we built the matrix $\mathbf{X}_{nxp}$ of the Cartesian coordinates of the Cα atoms in the family, where the first index labels the Cartesian coordinates of the core (n is the number of core positions), and the second one labels the structure. The covariance matrix is then obtained as $C_{ij} = < (x_i - < x_i >)(x_j - < x_j >) >$, where $<>$ means the average over all proteins. Through the spectral decomposition of $\mathbf{C}_{nxp}$, we get $\mathbf{C} = \mathbf{V}\Delta\mathbf{V}^{\mathrm{T}}$, where $\mathbf{V}$ is an orthogonal matrix containing the set of eigenvectors, or PCs, and Δ is a diagonal matrix containing the set of eigenvalues.

Expectation-maximization PCA (EM-PCA)

The standard PCA described earlier can only deal with gapless cores in which a Cα atom is present in all proteins and positions. We aim at dealing with the loose core of the protein family, defined as the set of positions in which at least a fraction of 2/3 of all the proteins have a Cα atom within 3 Å from the average of the family. To this task, we used a simplified version of the EM-PCA algorithm,[34] in which gap positions are treated as missing values. The procedure works like this. First, we apply the standard PCA to the structural alignment. Secondly, we construct the EM-PCA space based on the PCA eigenvectors. Suppose that $\mathbf{Y}_{mxp}$ is the coordinate matrix, which contains $\mathbf{X}_{nxp}$ and includes gap positions, and $\mathbf{A}_{kxp}$ is the passage matrix, obtained as $\mathbf{A}_{kxp} = (\mathbf{V}_{nxk}^{\mathrm{T}}\mathbf{V}_{nxk})^{-1}\mathbf{V}_{nxk}^{\mathrm{T}}\mathbf{X}_{nxp}$, where $m \geq n$ is the number of loose core coordinates and k is the number of PCA eigenvectors used. The passage matrix transfers the coordinates from the PCA space to the Cartesian space with the formula $\mathbf{X}_{nxp} = \mathbf{V}_{nxk}\mathbf{A}_{kxp}$. In the matrix $\mathbf{Y}_{mxp}$, the gapped position elements are initially approximated by the average values of the rest of the elements in the column, $<x_j>$. Then gapped eigenvectors can be obtained as $\mathbf{V}_{mxk} = \mathbf{X}_{mxp}\mathbf{A}_{kxp}^{\mathrm{T}}(\mathbf{A}_{kxp}\mathbf{A}_{kxp}^{\mathrm{T}})^{-1}$. In the usual implementation of the EM-PCA algorithm this procedure is iterated, deriving a new passage matrix. We do not do that here, since otherwise the elements of the strict core would change their values. Since the first passage matrix $\mathbf{A}_{kxp}$ is derived from the standard PCA, the coordinates of the strict core atoms are preserved in our procedure.

Normal modes calculations through the anisotropic network model (ANM)

The ANM[35] is a coarse-grained model, which models the folded state of the protein as a three-dimensional elastic network, where elastic forces are present between atoms in contact in the protein structure. The elastic constant is assumed to be the same for all contacts. The model has only one parameter, the contact distance d_0 at which two atoms are regarded as interacting, which was set to $d_0 =$ 15 Å. The potential energy of the protein (V) as a function of the displacement ($\mathbf{R}$) from the native conformation is thus $V = \mathbf{R}\mathbf{H}\mathbf{R}^{\mathrm{T}}/2$, where $\mathbf{H}$ coincides with the Hessian matrix of the second derivatives of the energy function. Diagonalization of $\mathbf{H}$ as $\mathbf{H} = \mathbf{U}\Lambda U^{T}$ yields $3N - 6$ NMs

contained in the eigenvector matrix **U**, N being the number of residues in the reference structure. The lowest frequency NMs corresponds to the directions of collective motion, along which the protein is most deformable. In the calculations presented here, we used as reference structure the average structure of a protein family.

Merging the EM-PCA and the ANM subspaces

The EVA sampling space is a combination of the EM-PCA subspace, which contains the available evolutionary information, and the ANM lowest frequency NMs, which contain the most relevant vibrational information and can complement limited evolutionary information. The evolutionary eigenvectors are only defined at positions in the loose core, whereas the vibrational eigenvectors are defined at all positions in the alignment. To join the two sets of eigenvectors at loose core positions, the NM are added one by one to the EM-PCA space, and orthogonalized with the Gram-Schmidt procedure.[36] Given the EM-PCA space $\mathbf{V}_{mxk}$ and a ANM eigenvector $\mathbf{t}_m$, the orthogonal part of the eigenvector to the space is obtained as $\mathbf{u}_m = \mathbf{t}_m - \mathbf{V}_{mxk}\,(\mathbf{V}^{\mathrm{T}}_{mxk}\mathbf{V}_{mxk})^{-1}(\mathbf{t}_m\mathbf{V}_{mxk})$. The space is then updated by adding $\mathbf{u}_m$ to the set.

Assessing the sampling subspace

Projection studies (superfamilies and CASP5)

The projection of the target structure on the EVA subspace was determined analytically for loose core atoms for which EVA coordinates are available, and its RMSD with respect to the target structure was calculated through a least squares fit.

Construction and assessment of the complete models for CASP5 targets

Starting from the projected structures, consisting only on the Cα trace of the core positions, all-atoms models were built. Backbone and side-chains were built with the programs MMTSB[37] and SCWRL,[38] respectively, while loops with less than six intervening residues where added with MODELER.[39] Finally, energy minimization with the Molecular Dynamics program AMBER8[40] was carried out to remove atomic clashes. We assessed the quality of the models with the program Procheck,[41] and calculated the RMSD with respect to the experimental structure, as well as the percentage of structure modeled, the percentage of correct χ1 and χ2 angles, the quality of the Ramachandran map, the dihedral angles and the number of bad contacts.

In addition, we also constructed FRAGBENCH models.[42] These are the best possible fragment-based model for each target obtained by combining fragments of structurally similar templates available at the time of the CASP5 experiment. The only change in the protocol with respect to the original article was the use of the latest version of the LGA software (03/2005[43]) to compute the GDT_TS score. FRAGBENCH models were compared with the best CASP5 models and with the best possible models in the EVA subspace.

Replica exchange monte carlo (REMC) simulations in the EVA subspace

Building the complete chain (loop modeling)

In the EVA subspace, coordinates are specified only for the Cα atoms in the loose core. Positions outside the core (loops) must be reconstructed using the core positions as attachment points. For these calculations, we represented the loops through the coordinates of their N and Cα atoms in the backbone. To model the loops, the bond lengths and angles were obtained from the AMBER 8 package. The loop building is initiated by randomly selecting dihedral angles starting from the N-terminal of the loop. Then the cyclic coordinate descent (CCD) algorithm[44] is employed for loop closure. The CCD method is a computationally fast and analytically simple method for adjusting the loop structure to connect its nearby core segments. The basic idea of CCD is to modify one torsion angle at a time in order to minimize the distance between the moving C-terminal residue and the target. The process is iterated until convergence. For the loops in which the N- or C- terminal are free, we randomly create the loop structure by giving random dihedral angles starting from the fixed end connected with the core residue, up to the free end.

Implementation of the REMC Method

REMC is a powerful tool for sampling in multiple dimensional phase spaces. In standard REMC, M replicas of the system are set at M different temperatures T_m ($m = 1,\ldots, M$). Then two steps are performed alternatively: first, each replica samples independently the phase space at its corresponding temperature for a certain number of steps. Secondly, the replicas at neighboring temperatures are exchanged according to an appropriate probability. Suppose that $X = \{x_m^{[i]}, x_n^{[j]}\}$ represents a state where the replica at T_m is in configuration $x^{[i]}$ and the replica at T_n is in configuration $x^{[j]}$. The exchange probability between $X = \{x_m^{[i]}, x_n^{[j]}\}$ and $X' = \{x_m^{[j]}, x_n^{[i]}\}$ is given by

$$w(X \to X') = \begin{cases} 1, & \text{for} \quad \Delta \le 0 \\ \exp(-\Delta), & \text{for} \quad \Delta > 0, \end{cases}$$

where $\Delta = (\frac{1}{T_n} - \frac{1}{T_m})(E_\mathrm{i} - E_\mathrm{j})$, and E_i and E_j are the energies of the i-th and j-th configurations, respectively. In this way, the simulation can overcome potential energy barriers at high temperature, and reach the local/global minima at low temperatures. The temperatures are distributed following the Okamoto method,[45] as $T_i = T_1(\frac{T_N}{T_1})^{(\frac{i-1}{N-1})}$, where T_1 and T_N are the highest and lowest temperatures, respectively, and N is the number of temperatures. Since the number of degrees of freedom of the EVA subspace is 50, we choose $N = 8$ replicas.

Core and loop positions are moved separately. The core part is sampled through REMC in the phase space constructed with EM-PCA and ANM eigenvectors, which describe the deviations from the average structure of the protein family. To make the sampling more efficient, we reduced the sampling space by adopting some limits. First, we measure the maximum deviation d_i from the average structure for the available structures along all directions i of the subsapce, and limit the sampling along each direction i in the range $[-3d_i, 3d_i]$. Second, since the structures of all members in the family are within 5 Å from the family average, we limit the sampling along each direction within 5 Å from the average. This is equivalent to limit the sampling in the range $[-5\sqrt{N_{\text{atom}}}, 5\sqrt{N_{\text{atom}}}]$ if all eigenvectors are orthonormalized, where N_{atom} is the number of loose core atoms.

For the loops, we used three kinds of moves. The first one is applied to the N- or C-terminal loop, having a free end. We randomly choose a residue i and then we randomly rotate the segment from residue i to the free end around the N_i-$C\alpha_i$ bond. The second kind of move is a local deformation that consists in rotating one randomly chosen N atom around the axis passing through its two nearby Cα atoms. The third kind of move is another local deformation, in which a segment of four residues in the loop is randomly chosen, for example (i, $i + 1$, $i + 2$, $i + 3$), and the positions of the atoms at the ($i + 3$)-th residue are slightly changed by varying the angle $N_{i+3} - C\alpha_{i+3} - N_{i+4}$ and rotating around the $C\alpha_{i+3} - N_{i+4}$ axis. Then the CCD algorithm described above is applied to adjust the segment (i, $i + 1$, $i + 2$, $i + 3$).

The starting point of the simulation was in all cases chosen setting the projections on the first two eigenvectors to [−3.34 Å, −3.34 Å]. This is 7.19 Å away from the target, whose projections on the same directions are [3.05 Å, 0.73 Å].

Simulated annealing (SA) and random search (RS)

We also test two additional stochastic minimization algorithms. The SA method[51] is very efficient in finding nearly optimal solutions of complex minimization problems. In this method, the proposed changes are accepted or discarded according to the Metropolis criterion at temperature T. Starting from high temperature, the temperature is decreased little by little, so that the simulation becomes more selective and tends towards some local minimum. The RS method[51] is not suitable for difficult minimization problems. In this method, the proposed moves are always accepted without any bias.

Evaluation of the sampling efficiency

The EVA subspace was first constructed using all proteins including the target to calculate the EM-PCA eigenvectors. In this way, the correct solution is present in the subspace. We then tested the efficiency of the minimization methods using designed energy functions, which include three terms:

$$E(\mathbf{c}) = k(\mathbf{c} - \mathbf{c}_1)^2 - \sum_{i=1}^{\text{nprot}} h_i \exp\left[-\frac{(\mathbf{c} - \mathbf{c}_i)^2}{2w_i^2}\right] + \sum_{i=1}^{\text{noise}} p_i \exp\left[-\frac{(\mathbf{c} - \beta_i)^2}{2\sigma_i^2}\right],$$

where $\mathbf{c}$ is the trial position in the EVA subspace, $\mathbf{c}_1$ represents the target structure and $\mathbf{c}_2 \ldots \mathbf{c}_P$ represent alternative structure in the family. The first term biases the energy function to have the global minimum with funnel-like shape centered in the target. The parameter $k = 0.001$ controls the steepness of the funnel, and it is set to a small value, for the purpose of making the sampling more challenging. The second term is a superposition of Gaussian functions, each centered at the i-th protein structure of the family, with coordinates $\mathbf{c}_i$. The parameters h_1 and w_1, corresponding to the depth and the width of i-th Gaussian peak, are both positive. The role of this term is to simulate spurious attractors corresponding to real structures in the family. In the third term, random noise is generated through a superposition of Gaussian functions, where noise is the number of peaks. Except for h_1, which was set to $h_1 = 5.0$, all other parameters were randomly chosen in a range of [0.8 p_m, p_m], where p_m is the maximum limit, which was 1.25 for the h_i, 4.2 and 4.5 for the w_i different from $i = 1$, 8.4 and 9.0 for w_1, 3.5 and 4.5 for the p_i, and 8.4 and 9.0 for the σ_i, depending on which kind of noise density was used (see below). Noise peaks are designed to be regularly distributed in the phase space, and their height parameters p_i for neighboring Gaussian peaks have opposite signs, so that barriers that may increase the difficulty of the sampling surround all energy minima.

E_{gap}, the difference between h_1 and the maximum value of p_i, is used in Table III to parameterize the ruggedness of the designed energy landscape. Another related measure of ruggedness was obtained by calculating the Z-score, defined as follows:

$$Z\text{-score} = \frac{E_{\text{min}} - \langle E \rangle}{\sigma}$$

where E_{min} is the global energy minimum, $\langle E \rangle$ and σ are the average and standard deviation of the energy function in the complete landscape. In this article, we designed four landscapes having Z-score $= -3.437$, -3.395, -3.267, and -3.221, ranked for increasing ruggedness. In all cases, we only changed the parameters p_m, keeping the other parameters fixed. The temperatures are distributed as (2.000, 0.636, 0.313, 0.186, 0.123, 0.087, 0.065, 0.050).

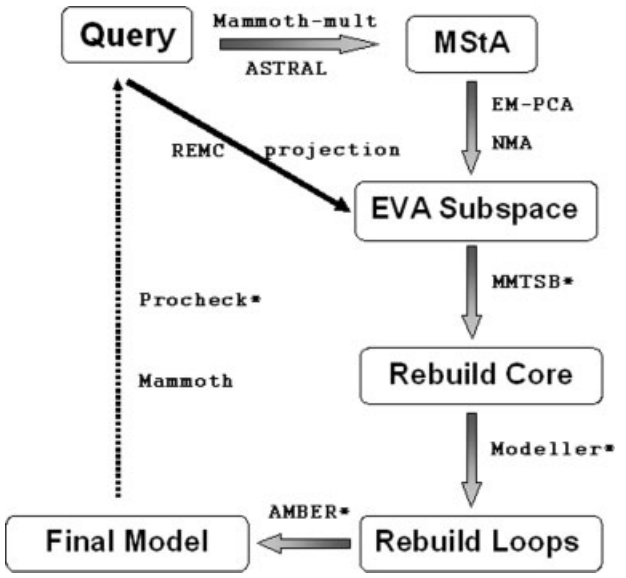

Figure 1

Integration of the different steps and programs involved in computational experiments 1 and 2 (see Methods for more details). Programs labeled with asterisks correspond to third party programs.

RESULTS

This section is organized as follows: first we introduce the results obtained with the EM-PCA algorithm designed to maximize the evolutionary core size in the EVA subspace. Secondly, we present the results that lead us to the definition of the EVA subspace using the evolutionary core previously found. And finally, we present our evaluation of the sampling properties of the REMC algorithm in the EVA subspace. The first two sections can be integrated in the general scheme to build and evaluate our protein models shown in Figure 1. In this figure, the different software tools employed are also highlighted (those labeled with an asterisk correspond to third party software).

Enlarged support of the PC vectors

The standard PC analysis can only be applied at strict core positions that are present in all proteins of the family. To enlarge the support of the PC vectors, we adopted Expectation Maximization PCA[34] (EM-PCA). With this technique, it is possible to deal with positions with gaps in the alignment, which are treated as unknown values. We applied the EM-PCA analysis to the loose core of the protein family, that is, all positions where a residue is present in at least two thirds of the proteins within 3 Å from the average of the family. EM-PCA greatly increases the scope of standard PCA.[46] The core size distribution for all 547 proteins using both standard PCA and EM-PCA shows large differences (Fig. 1 of the Supplementary Material). The loose cores adopted in EM-PCA are much larger than the loose cores needed for standard PCA, with the mode of the core size increasing from 60% for the strict core to 80% for the loose core. Whereas only 22.5% of the proteins have strict core size larger than 60% of the structure, this percentage reaches 69% for loose cores.

Accuracy of the reduced subspaces

We next tested the geometric accuracy of the reduced subspaces obtained by combining the evolutionary PCs and the lowest frequency NMs. We compared the RMSD between the target (native structure) and its projection on the subspace (*RMSD_project*) to the RMSD between the target and the template with largest sequence identity (*RMSD_closest*). The projection is the best model that can be found in the subspace, so that *RMSD_project* measures the geometric accuracy of the sampling subspace. This is compared with the accuracy *RMSD_closest* of the best model that can be obtained from a standard homology modeling approach, in which the protein with the largest sequence identity is aligned with the target using the correct structural alignment.

These data are shown in Figure 2. Figure 2(A) refers to the subspace that was used by Qian *et al.*,[27] defined by the three largest PCs, Figure 2(B) refers the subspace defined by all PCs. One can notice that the first subspace, although performing better than the best possible standard homology model, is significantly less accurate than using all the available PCs. For the 547 proteins in this study, the average value of *RMSD_project* is 1.40 Å using only the first three PCs, and it decreases to 1.09 Å using all PCs.

To further improve this result, we combined the PCs of the evolutionary variation with the lowest frequency NM calculated with the anisotropic network model[35] (ANM), using the average structure of the family as a reference. Previously, Leo-Macias *et al.* demonstrated that the subspaces built with the PCs and with the low-frequency NMs overlap significantly. Thus low-frequency NMs can complement the PCA subspace as a surrogate of the unknown evolutionary information not present in the multiple structure alignment because of poor structural sampling. We show in Figure 2(C), the subspace obtained with the 50 NMs of lowest frequency, and in Figure 2(D) all PCs plus NMs up to 50 dimensions. It can be seen that the projections in all subspaces are on the average closer to the target than the most similar structure in the family. It can also be seen that the PC eigenvectors are more accurate to predict unknown evolutionary variance than the NMs, since the average accuracy obtained with 50 NMs is comparable to that obtained with only three PCs. This agrees with our previous observations, and indicates that the PC eigenvectors

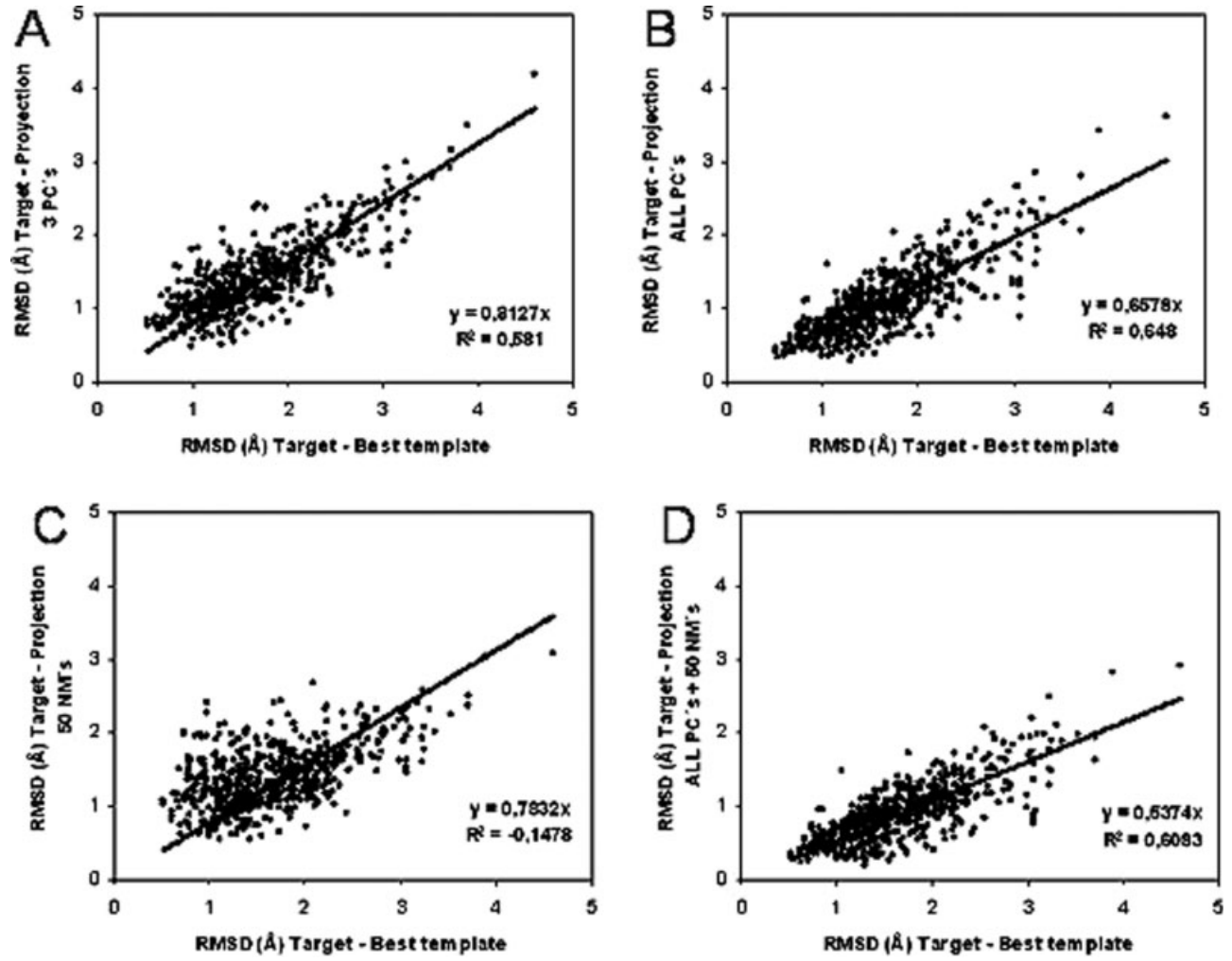

Figure 2

*RMSD of the target structure to its closest homologue (horizontal axis) versus the RMSD of the target structure to the one projected in the reduced subspace (vertical axis). Only the loose core is considered. The different subspaces are obtained with (**A**) the three main EM-PCA eigenvectors; (**B**) All available eigenvectors (**C**) The 50 lowest frequency ANM normal modes (**D**) All the EM-PCA eigenvectors plus the lowest frequency ANM normal modes up to 50 dimensions.*

are much more valuable than the NMs for constructing the sampling subspace. We quantified the average improvement obtained with the best model in a sampling subspace with respect to the best standard homology by measuring the slope of the linear regression between *RMS_project* and *RMS_closest*. The smaller this slope, the more accurate the sampling subspace is. A slope of 1 means that there is no improvement with respect to homology modeling on the average, and a slope of 0 corresponds to the complete space. This is shown in Figure 3 versus the number of dimensions for the subspaces obtained by adding the lowest NMs to the entire set of PCs. Obviously, the larger the number of dimensions, the more accurate the subspace. In the range of values studied, the average improvement for adding one additional NM is 0.4%, and we would reach a slope of zero by

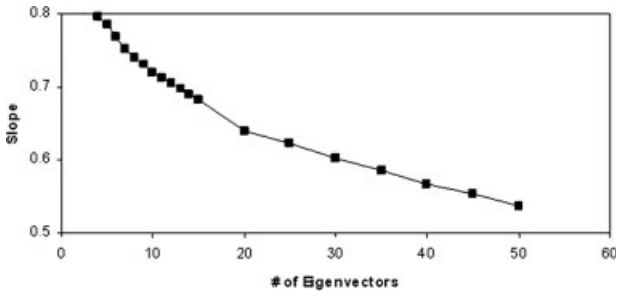

Figure 3

Slopes of the linear fits corresponding to the different subspaces (i.e. the slopes plotted in Figure 2 for a few representative cases) versus the number of dimensions used to create the reduced subspace. The smaller the slope, the higher the subspace accuracy.

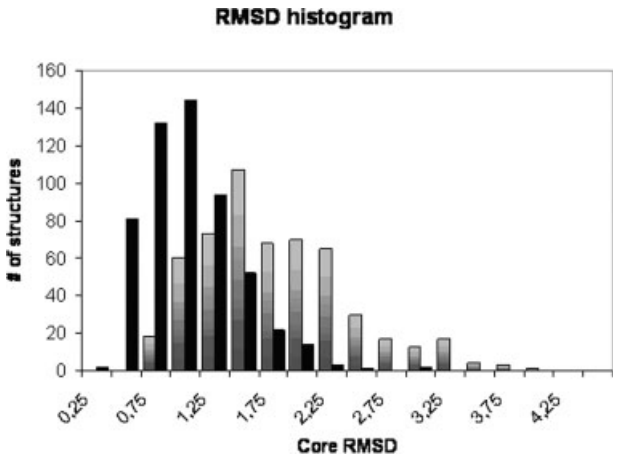

Figure 4

Frequency distribution for the RMSD between target and projection (black) using the EM-PCA + ANM 50 dimensional space (see main text for details), and for the most similar structure in the family (grey).

using all NMs, which is equivalent to use all the degrees of freedom of the backbone. There is however a trade-off with the volume of conformational space. Consideration of the computational requirements suggests that the maximum number of dimensions tractable with the REMC method is of the order of 50, and for this number of dimensions the accuracy is rather good, with an average improvement of 46% between the best possible model in the subspace and the best possible homology model. We also show in Figure 2(C) the results when employing only NMs to define the EVA subspace. The lack of correlation was expected, as the displacement along the lowest frequency modes only provides a fraction of evolutionary information.[35] The purpose of Figure 2(C) is to show that refinement methods based only in NM analysis (and, since low-frequency NMs significantly overlap with the essential dynamics eigenvectors derived from molecular dynamics simulations, most likely from molecular dynamics simulations as well), are unlikely to be successful in the refinement problem, unless the full spectrum is introduced. But they can be useful to complement the evolutionary eigenvectors, as shown in Figure 2(D).

To summarize our results, in Figure 4 we show an histogram of the maximum attainable accuracies in our sampling subspace and a similar histogram of the RMSD computed with the most similar structure in the family (as a surrogate of the best attainable homology model). The shift between both distributions is evident. With our sampling subspace, about 65.6%, of the proteins have native-like structures with RMSD smaller than 1 Å, and 92.3% for the RMSD smaller than 1.5 Å. From now on, we will refer to the reduced subspace, consisting of all PCs and NMs up to 50 dimensions, with the acronym evolutionary and vibrational armonics (EVA) subspace. We notice that the coordinates of the EVA subspace can only be defined at positions that belong to the loose core of the protein, where the EM-PC analysis of the evolutionary variation can be performed.

Next, we also evaluated the EVA sampling subspace using a set of homology modeling targets of the previous CASP5 experiment. We compared the potential accuracy of the best model in the EVA subspace with the accuracy of the models submitted to the CASP experiment, which were generated through state of the art homology modeling protocols, and with the potential accuracy of the Fragbench approach.[29] In this approach, for each target structure one builds the best model that can be obtained using as templates fragments of known PDB structures, combined in the optimal way. These are the best possible models that can be obtained using known structures, even in the absence of a global structural similarity.

We used the set of targets as divided in domains by CASP5 organizers, excluding the target T0186_3, for which we did not find any significantly similar structure in our library of templates (see Methods Section). This same set was already used to test the FRAGBENCH approach. We projected each target structure on its corresponding EVA subspace, and measured the RMSD between the projection and the target. For comparison, we also measured the RMSD between the target and the best available template. In Figure 5 one can see that ~35% of the targets can be represented on their EVA subspaces with RMSD $<$ 1 Å, and 60% of them with RMSD $<$ 1.5 Å, whereas the percentage of targets with RMSD $<$ 1.5 Å from their best template is lower than 20%. This confirms that the reduced subspace is potentially rather more accurate than the best possible standard homology modeling approach.

We also measured GDT_TS scores to compare the accuracy of the projections on the EVA subspace to that of the best models in CASP5, and of the best models that can be generated combining fragments with FRAGBENCH (see Methods Section). FragBench attempts, given a target-template structural alignment, to build the best possible geometrical model by scanning the whole protein structure database (excluding the target) for suitable structural fragments with high structural similarity to the target structure (used as query), which are sequentially added to build the final model. Thus, FragBench provides one realization of the best possible model that, in terms of geometrical features, can be attained with the knowledge of the complete PDB, using the structural information of the target as a query and a "perfect" (structural) alignment. In this way, FragBench provides an upper bound estimate of the best possible model that can be expected. When testing the EVA subspace, the best we can hope for is to get to the FragBench line. In contrast, when we test the geometrical accuracy of the EVA subspace, neither the target structure nor the struc-

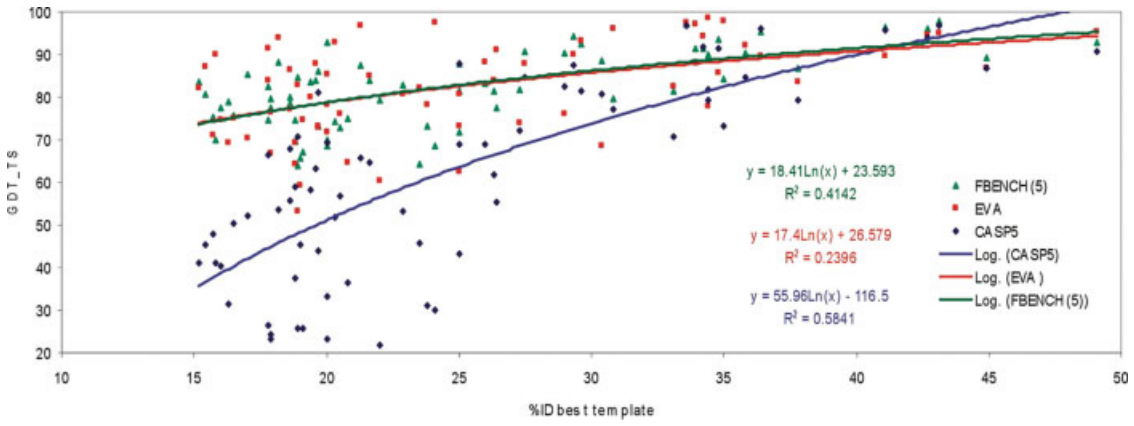

Figure 5

GDT_TS results for the CASP5 targets. We compare the results obtained in the CASP5 experiment (blue) with the best possible result in the EVA space (pink) and with the best possible result obtainable combining fragments of known structures (turquese).

ture database is employed. When we characterize the EVA subspace, the structure closest to the target is mathematically deduced from a linear combination of the structures of the known homologues. At the other extreme, we also compared in Figure 6 these results with the best possible model obtained by the CASP5 participants. We considered these data as our lower bound. Models with our method should be at least as good as those produced by the standard state of the art. The results of these comparisons are shown in Figure 6. For sequence identity below 40%, the projections on the EVA subspace are as accurate as the FRAGBENCH models, indicating that EVA projections reach the maximum accuracy that can be attained using known protein structures, and they are much more accurate than the best models evaluated in CASP5. For larger sequence identity, the best models in CASP5 are as accurate as those that can be obtained in the EVA subspace and in the FRAGBENCH approach, and of quality comparable to low-resolution experimental structures. In our opinion, this can be explained as a result of the low dimensional, topological nature of the structural transitions taking place during evolution.[26]

So far, we only considered the Cα trace of the protein structure. We next generated all atom models for each projected structure. Finally, we minimized the energy of each model with the AMBER program,[40] in order to remove atomic clashes. We studied the quality of the model before and after minimization by assessing the RMSD, the statistics of the Ramachandran plot, the $\chi 1$ and $\chi 2$ angles and the number of bad contacts. The results are presented in Table III (before minimization) and Table IV (after minimization) (Supplementary Material). After the minimization the RMSD typically increased for the residues in the core, which were represented in the reduced subspace, but not for the loops. The bad contacts nearly disappeared and the number of residues in the most favored regions of the Ramachandran map increased, while χ angle values were not affected. This indicates that the projection of the target on the EVA subspace has some unphysical features, but those can be easily removed through energy minimization.

Figure 7 shows some examples of superimposition of the target structures with the model built in the EVA subspace and with the best template available at the time when CASP5 took place. It can be seen that the best projection in the subspace improved considerably with respect to the best template.

Sampling algorithm

We now turn to the problem of sampling low-energy conformations in the EVA subspace. For that, we have studied the applicability and efficiency of the REMC[45,47–49] optimization method.

In our first test, we used as energy function simply the RMSD between the sampled structure and the target. This smooth energy function does not pose a challenge to the minimization algorithm, but was chosen only for validation purposes. We model the complete chain, consisting of the protein core and the nonconserved parts (loops). The core is allowed to move only in the reduced EVA subspace, while the motion of the loops is not restricted, and it is performed through the CCD algorithm[44,50] (see Methods Section). The structure selected as a target is excluded in the computation of the PC. We compared the core RMSD obtained by REMC simulations with the analytical minimum of the energy function, which is obtained by projection (Fig. 2 of the Supplementary Material). The two values agree in all cases within 0.025 Å, which means that in all cases the REMC

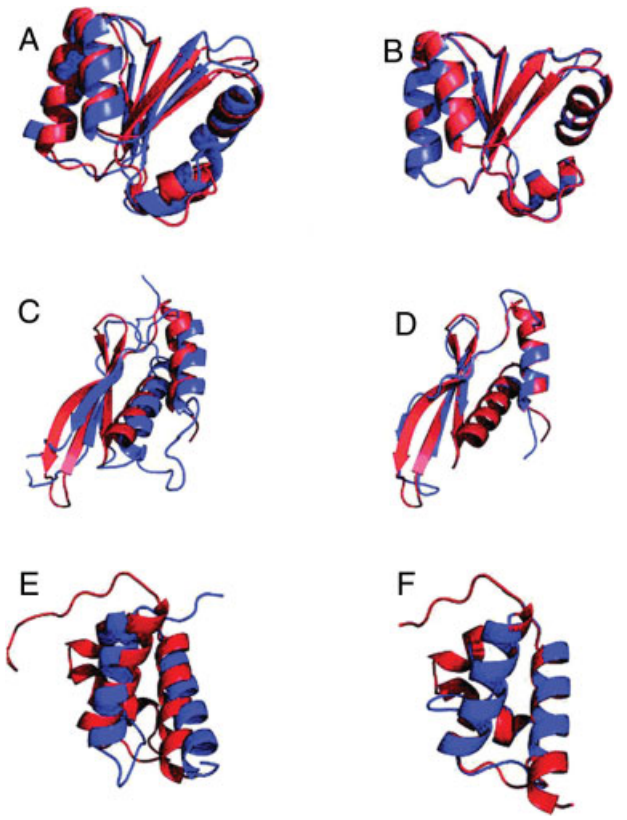

Figure 6

*Structural superimpositions of CASP5 targets with the best homology model obtained in the CASP5 competition (i.e., most similar structure according to MAMMOTH) (**A,C,E**), and with the structure reconstructed from projection on the EVA space using the 50-dimensional space (**B,D,F**). Targets are in blue, models in red. (A) T0150-cSTR ($RMSD_{core}$ = 2.16 Å); (B) T0150- Model ($RMSD_{core}$ = 0.39 Å); (C) T0184_2-cSTR ($RMSD_{core}$ = 2.33 Å); (D) T0184_2-Model ($RMSD_{core}$ = 0.33 Å; (E) T0170-cSTR ($RMSD_{core}$ = 3.65 Å); (F) T0170-Model ($RMSD_{core}$ = 0.11 Å).*

algorithm gets quite close to the global minimum. The accuracy for the loops is larger than the one for the core, since the loops were not restricted to the EVA subspace. Loops were sampled separately from the core, and also for them the REMC algorithm was always able to get very close to the global minimum. The combined accuracy of the core and the loops was below 1 Å for almost 80% of the proteins (not shown).

Next, we tested the method using more challenging energy functions, with multiple minima and various degrees of ruggedness. These energy functions were defined in such a way that their global minimum corresponds to a neighborhood of the projection of the target structure in the EVA subspace, but many false minima are present as well. The ruggedness of the energy function is parameterized through the energy gap and the resulting *Z*-score. A smaller energy gap, yielding a less negative *Z*-score, corresponds to a more rugged landscape. Figure 8 illustrates, as an example, the landscape constructed using 2120 noise peaks, with energy gap set to $E_{gap} = 0.5$, which corresponds to *Z*-score = −3.437. A 2-D cross-section of the energy landscape, corresponding to the first two PC, is represented in Figure 8(A). This cross-section represents typical characteristics of the whole landscape. In the right hand side plot of Figure 8(A), we represent a typical low-temperature trajectory of the REMC simulations. It can be seen that the sampling is very efficient. It quickly gets close to the target, and then remains in its attraction basin. Figure 8(B) shows the histogram of the energy, which is a Gaussian-like peak centered near zero. Figure 8(C) shows the canonical probability distributions at eight temperatures. It can be seen that there is enough overlap between neighboring distributions, which is a necessary condition for replica exchange to work properly. Through replica exchanges, the sampling can overcome the energy barriers at high temperature, it avoids to get trapped into the local minima at low temperature, and it can reach the global minimum.

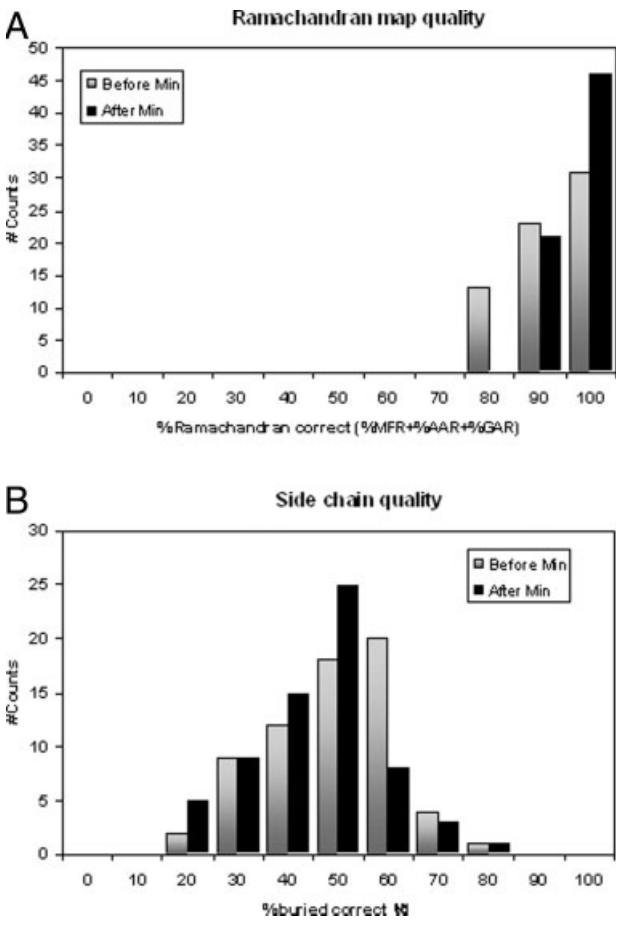

Figure 7

***A:** Distribution of the % correct residues in the Ramachandran map for the EVA-based models in the CASP5 dataset before (black) and after (grey) energy minimization. **B:** Similar to A, but with reference to the quality of side chain packing (using a Δ = 30° and a burial cutoff of 30%).*

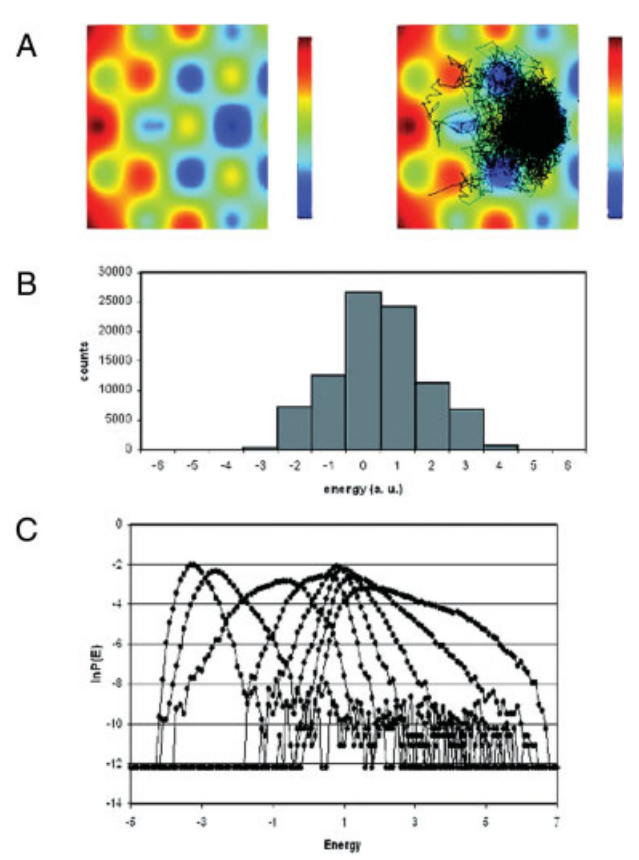

Figure 8

*Example of REMC sampling assessment: (**A**) 2D projection of the energy landscape and evolution of the lowest temperature simulation; (**B**) Energy spectrum of the energy landscape; (**C**) Energy distribution covered by the REMC simulation.*

To show that the energy landscapes that we studied are truly challenging, we also sampled them through simulated annealing[51] (SA), an efficient method for energy optimization in landscapes with multiple local minima, although not as much as the REMC method, and also through random search[51] (RS), which is not expected to be an efficient method for such a difficult task. The REMC sampling parameters are the same as those used in the example in Figure 8. The sampling steps for SA and RS are set to be 100,000, the same number as in the REMC, and the temperature for SA is decreased from 2.0 to 0.05, the same range as for REMC simulations.

We studied the performances of REMC sampling, compared to SA in landscapes with different levels of ruggedness. Results are reported in Table I, with standard deviations shown in parentheses. We considered four landscapes, obtained by combining two different distributions of noise peaks (2120 noise peaks at a distance of 2.5 Å between nearby peaks, and 2903 noise peaks at a distance of 2.4 Å between nearby peaks) with two values of the energy gap, $E_{gap} = 0.5$ au and $E_{gap} = 1.5$ au. The corresponding Z-scores are reported in the second row of Table I. For each of the four landscapes, we performed 10 simulations with each method. The third row reports the global minimum of the energy, which can be calculated analytically from the definition of the energy function. We then report E_{min}, the average of the lowest energy sampled in the 10 simulations, and $rmsd_{min}$, the average of the RMSD corresponding to the energy minima. From the table, we can see that the REMC simulations always get very close to the target, with a minimum energy at most within 15% of the global minimum. This means that the simulation always reaches the attraction basin of the global minimum. The $rmsd_{min}$ is on the average smaller than 0.4 Å, which means that the model found through the EVA-REMC method is rather accurate. By contrast, the energy minimum found with the SA method indicates that the simulations have been trapped in some local minima, corresponding to a minimum RMSD close to 1 Å.

DISCUSSION

In this article, we have presented a sampling strategy to refine protein models built by homology. In its development, we first, separated the problem of defining a proper

Table 1

Summary of the REMC Sampling Results

		$N_{peaks\_noise} = 2120$		$N_{peaks\_noise} = 2903$	
	Z-score	−3.437	−3.221	−3.395	−3.267
	E_{glob}	−4.765	−4.817	−4.778	−4.828
REMC	$<E_{min}>$	−4.199 (0.056)	−4.225 (0.063)	−4.183 (0.058)	−4.225 (0.064)
	$<rmsd\ E_{min}>$	0.381 (0.021)	0.387 (0.021)	0.366 (0.028)	0.383 (0.027)
SA	$<E_{min}>$	−1.621 (0.251)	−1.690 (0.187)	−1.523 (0.416)	−1.758 (0.141)
	$<rmsd\ E_{min}>$	1.047 (0.077)	1.031 (0.067)	1.032 (0.114)	1.011 (0.042)
RS	$<E_{min}>$	0.156 (0.160)	0.104 (0.159)	0.243 (0.094)	0.147 (0.124)
	$<rmsd\ E_{min}>$	1.815 (0.133)	1.775 (0.118)	1.813 (0.108)	1.818 (0.098)

Results for REMC (Replica Exchange Monte Carlo), SA (Simulated Annealing) and RS (Random Search) correspond to averages over 10 simulations. Average and standard deviation (in parenthesis) values are shown. See text for more details.

free energy function, which will be treated in a forthcoming publication, and focus our attention into the task of energy optimization. Our philosophy is that sampling is a major concern in homology modeling that must be studied in its own right, in controlled situations where the global minima can be defined, roughness of the configurational space controlled, and a quantitative measure of conformational sampling defined. However, in these conditions we need to pay the toll of employing an artificial energy function, as we have done in this article, implying that our conclusions may not directly translate to a "real" situation. In our view, however, whether or not our results can be extrapolated depends mainly on whether our conditions conform or not with the features of the energy landscape available when employing *a* "real" function. We think it does, for two reasons. First, we have introduced in our experiments the level of roughness to be expected with real energy functions. For example, Zhang and Skolnick[52,53] studied the *Z*-score of various typical energy functions employed in structure prediction at that time, and compared them to the *Z*-scores expected for real proteins using data extracted from calorimetric measurements. According to these authors, typical *Z*-scores for the energy functions available at that time were between 4 and 15, still a far cry from the *Z*-score employed by Nature. We have employed *Z*-scores in the order of 3–4 in our simulations (see Table I). This means that we have employed, in our simulations, comparable or even harsher conditions than those expected with the realistic energy functions available at that time. Second, we have preliminary data with realistic energy functions indicating that the method also performs very well in real situations. We will report these data in a forthcoming publication (RH and ARO, unpublished data).

Thus, the challenge we address here is how to devise a robust dimensionality reduction algorithm, able to condense as much evolutionary information as possible, so that powerful optimization methods such as REMC can be applied to thoroughly sample conformational space. REMC can only work if there is substantial overlap between the regions of phase space corresponding to different temperatures. And to ensure this condition, the number of temperatures to be simultaneously simulated must increase as the square root of the number of dimensions.[47] Therefore, REMC becomes computationally unfeasible for realistic protein models, such as those required in homology modeling in normal conditions. However, the two competing requirements of small dimensionality and accurate representation are satisfied in the EVA subspace. This subspace combines evolutionary and topological information, to obtain a reduced set of directions along which structural divergence is most likely to occur during evolution. The EVA subspace can be defined for each superfamily of homologous proteins. It consists of 50 dimensions, by far less than the number of degrees of freedom of a medium size protein chain. We have shown here that its accuracy is large enough so that all proteins in a superfamily can be represented within 1 Å accuracy. When assessing the two contributions to be EVA subspace, it is clear from the results presented in Figure 2 that the evolutionary eigenvectors are much more valuable than the vibrational eigenvectors to predict unknown evolutionary variation within a superfamily. In our view, this is due to the fact that the evolutionary eigenvectors capture both neutral and functional changes in protein structures, whereas the vibrational modes of low frequency define the directions along which a protein structure is more deformable and neutral structural variation is more likely to happen. Functional changes are expected to be the biggest changes in protein evolution, and the most difficult to model through homology.

Comparing the best model present in the EVA subspace with the best model that could be obtained through standard comparative modeling (i.e., using the best possible template structure and the correct structural alignment), we see that the RMSD with the target decreases by almost 50% on average (see Figs. 2 and 4). The results are a considerable improvement with respect to the results previously obtained by Qian *et al.*,[27] who used only the three largest PCs of the evolutionary variation. According to our calculations, the best models in this subspace decrease the RMSD to the target structure with respect to the optimal homology model by 19% on the average, but they are significantly less accurate than models in the EVA subspace. Another important technical improvement presented here with respect to the work of Qian *et al.* is that the portion of the protein that can be modeled in the reduced subspace has been greatly enlarged. Whereas normal PCA can only be applied at positions present in all proteins of the superfamily, we adopted Expectation-Maximization PCA, which treats gaps in the alignment as unknown values. In this way, we could apply the method to the loose cores of the superfamilies, that is all positions with not less than 66% conservation.

In the second part of this article, we tested the REMC sampling method in the EVA subspace. First, we used the RMSD with the target as an energy function. Simulations always got very close to the analytically known global minimum. A more challenging test was then performed with designed energy landscapes of variable roughness. Again, the analytically known global minimum coincided with the projected structure, and we added to it regularly distributed peaks of Gaussian noise, and terms that create additional local minima corresponding to the other proteins of the superfamily. The REMC method was able to get very close to the global minimum in all cases. The minimum energy attained was only 15% higher than the global minimum, corresponding to structures less than 0.4 Å far away from the target (Table III). By contrast, the simulated annealing method, which is also an efficient optimization algorithm, in a comparable computa-

tion time did not get lower than structures with energy 75% higher than the global minimum, corresponding to 1 Å RMSD from the target structure.

In summary, the results presented here offer a promising avenue to address the problem of homology model refinement, through the definition of a high accuracy and low dimensionality sampling space from the observed evolutionary and vibrational variation of the protein family, and the use of a very efficient optimization method. A β version of the code is available for interested academic laboratories.

ACKNOWLEDGMENTS

Generous allocation of computer time at the Barcelona Supercomputer Center is gratefully acknowledged. ALM acknowledges a predoctoral FPI fellowship from the Education and Science Ministry of Spain. BCM is funded by a postdoctoral fellowship from Universidad Nacional Autónoma de México.

REFERENCES

1. Chandonia JM. Brenner SE. The impact of structural genomics: expectations and outcomes. Science 2006;311:347–351.
2. Fiser A, Feig M, Brooks CL, IIIrd, Sali A. Evolution and physics in comparative protein structure modeling. Acc Chem Res 2002; 35:413–421.
3. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
4. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol 2006;16:172–177.
5. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A. Protein structure modeling for structural genomics. Nat Struct Biol 2000;7(Suppl):986–990.
6. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642.
7. Topf M, Sali A. Combining electron microscopy and comparative protein structure modeling. Curr Opin Struct Biol 2005;15:578–585.
8. Tramontano A. Worth the effort. An account of the Seventh Meeting of the Worldwide Critical Assessment of Techniques for Protein Structure Prediction. FEBS J 2007;274:1651–1654.
9. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
10. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round 6. Proteins 2005;61(Suppl 7):3–7.
11. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53 (Suppl 6):352–368.
12. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;45(Suppl 5):163–170.
13. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 2003;53 (Suppl 6):334–339.
14. Valencia A. Protein refinement: a new challenge for CASP in its 10th anniversary. Bioinformatics 2005;21:277.
15. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61(Suppl 7):27–45.
16. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 2003;46:2895–2907.
17. Ortiz AR, Gomez-Puertas P, Leo-Macias A, Lopez-Romero P, Lopez-Vinas E, Morreale A, Murcia M, Wang K. Computational approaches to model ligand selectivity in drug design. Curr Top Med Chem 2006;6:41–55.
18. Giorgetti A, Raimondo D, Miele AE, Tramontano A. Evaluating the usefulness of protein structure models for molecular replacement. Bioinformatics 2005;21 (Suppl_2):ii72–ii76.
19. Claude JB, Suhre K, Notredame C, Claverie JM, Abergel C. CaspR: a web server for automated molecular replacement using homology modeling. Nucleic Acids Res 2004;32(Web Server issue):W606–W609.
20. Read RJ. Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D Biol Crystallogr 2001; 57:1373–1382.
21. Chakravarty S, Sanchez R. Systematic analysis of added-value in simple comparative models of protein structure. Structure 2004; 12:1461–1470.
22. Chakravarty S, Wang L, Sanchez R. Accuracy of structure-derived properties in simple comparative models of protein structures. Nucleic Acids Res 2005;33:244–259.
23. Chen J, Brooks CL, IIIrd. Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 2007;67:922–930.
24. Lu H, Skolnick J. Application of statistical potentials to protein structure refinement from low resolution ab initio models. Biopolymers 2003;70:575–584.
25. Misura KM, Baker D. Progress and challenges in high-resolution refinement of protein structure models. Proteins 2005;59:15–29.
26. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophys J 2005;88:1291–1299.
27. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc Natl Acad Sci USA 2004; 101:15346–15351.
28. Lupyan D, Leo-Macías A, Ortiz AR. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 2005;21:3255–3263.
29. Mitsutake A, Sugita Y, Okamoto Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 2001;60:96–123.
30. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32(Database issue):D189–192.
31. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
32. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
33. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
34. Skocaj D, Bischof H, Leonardis A, A Robust. PCA algorithm for building representations for panoramic images. Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Proceedings, Part IV. Vol 2353, Copenhagen, Denmark; 2002. pp 761–775.
35. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001;80:505–515.
36. Arfken G, Weber HS. Gram-Schmidt orthogonalization. In: Press A, editor. Mathematical methods for physicists. Orlando; 1985. Academic Press, pp 516–520.
37. Feig M, Karanicolas J, Brooks CL, IIIrd.MMTSB Tool Set. MMTSB NIH Research Resource, The Scripps Research Institute; 2001.

38. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.
39. Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. Bioinformatics 2003;19:2500–2501.
40. Case DA, Cheathan TE, IIIrd, Darden T, Gohlke H, Luo R, Merz KM, Jr, Onufriev A, Simmerling C, Wang B, Woods R. The Amber biomolecular simulation programs. Comput Chem 2005;26:1668–1688.
41. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck—a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
42. Contreras-Moreira B, Ezkurdia I, Tress ML, Valencia A. Empirical limits for template-based protein structure prediction: the CASP5 example. FEBS Lett 2005;579:1203–1207.
43. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
44. Canutescu AA, Dunbrack RL. Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 2003;12:963–972.
45. Mitsutake A, Sugita Y, Okamoto Y. Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. II. Application to a more complex system. J Chem Phys 2003;118:6676–6688.
46. Gonzalez RC, Woods RC. Digital image processing. Reading, MA: Addison-Wesley; 1992.
47. Hukushima K, Nemoto K. Exchange Monte Carlo method and application to spin glass simulations. J Phys Soc Jpn 1996;65:1604–1608.
48. Hukushima K, Takayama H, Nemoto K. Application of an extended ensemble method to spin glasses. Int J Mod Phys C: Phys Comput 1996;7:337–344.
49. Mitsutake A, Sugita Y, Okamoto Y. Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test. J Chem Phys 2003;118:6664–6675.
50. Sharma G, Badescu M, Dubey A, Mavroidis C, Tomassone SM, Yarmush ML. Kinematics and workspace analysis of protein based nano-actuators. J Mech Design 2005;127:718–727.
51. Press WH, Press WH. Numerical recipes in FORTRAN: the art of scientific computing. Cambridge: Cambridge University Press, 1992; xxvi,963 p.
52. Zhang L, Skolnick J. What should the Z-score of native protein structures be? Protein Sci 1998;7:1201–1207.
53. Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.