Identification of protein structural neighbors to a query is fundamental in structure and function prediction. Here we present BS-align, a systematic method to retrieve backbone string neighbors from primary sequences as templates for protein modeling. The backbone conformation of a protein is represented by the backbone string, as defined in Ramachandran space. The backbone string of a query can be accurately predicted by two innovative technologies: a knowledge-driven sequence alignment and encoding of a backbone string element profile. Then, the predicted backbone string is employed to align against a backbone string database and retrieve a set of backbone string neighbors. The backbone string neighbors were shown to be close to native structures of query proteins. BS-align was successfully employed to predict models of 10 membrane proteins with lengths ranging between 229 and 595 residues, and whose high-resolution structural determinations were difficult to elucidate both by experiment and prediction. The obtained TM-scores and root mean square deviations of the models confirmed that the models based on the backbone string neighbors retrieved by the BS-align were very close to the native membrane structures although the query and the neighbor shared a very low sequence identity. The backbone string system represents a new road for the prediction of protein structure from sequence, and suggests that the similarity of the backbone string would be more informative than describing a protein as belonging to a fold. Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.016808, 1–7, 2012.

Determining the structures of membrane proteins remains a relatively unexplored frontier in structural biology (1). Current computational methods include de novo protein modeling and comparative modeling. Compared with de novo modeling, the comparative modeling is more successful when sequence homologies are available. However, because relatively few membrane proteins have been identified through experimentation, building membrane protein conformations remain an extremely difficult and daunting undertaking.

In comparative modeling or other methods based on known protein structures, the identification of the best structure neighbor (template), if indeed any are available, is critical. The typical method of template identification relies on serial pairwise sequence alignments aided by database search engines such as FASTA (2) and BLAST (3). More sensitive methods based on multiple sequence alignments, including PSI-BLAST (4), CLUSTALW (5), and HMMER (6), are available. MSAs have been shown to produce a greater number of potential templates and to better identify templates for sequences that have homologue relationships to other solved structures. However, when there is no significant homology found, most of the target-template pairs are evolutionarily too distant to be detected with the current threading approaches (7). If no other information about the target is known, aside from the sequence, it becomes difficult to identify possible templates and the correct threading of the target onto a known structure (it is an NP-hard problem for some models of threading (8)). On the other hand, when the query structure is known, structural alignment tools such as Structal (9), CE (10), SSM (11), TM-Align (12), and FragBag (13) can usually retrieve structural neighbors quickly and accurately, including proteins that share a low sequence similarity. From a structural view, the current protein data bank (PDB)1 may be the best approach to solve the problem of protein structure prediction (14). Exploiting a novel strategy between sequence-based and structure-based methods for retrieving protein neighbors...
may greatly change the landscape of protein structure prediction.

Here, we introduce the BS-align algorithm, a backbone string-based pipeline identifying structural neighbors from primary sequence by representing a protein structure as backbone string, which is defined on detached regions in Ramachandran space (15). This predicted representation is then used to search a candidate set of protein structural neighbors by backbone string alignment, and to show that very close neighbors of membrane proteins can be identified from their sequences.

First, we developed an accurate predictor of backbone strings, based on two innovative technologies: a knowledge-driven sequence alignment and a backbone string element profile embodying structural evolution information. Then, we retrieved backbone string neighbors by aligning the predicted backbone string of a query sequence against a backbone string database composed of known proteins. The retrieved top \( n \) sequences constitute a candidate set of neighbors. By testing a benchmark set of protein sequences, our approach outperformed most threading methods. Finally, we demonstrated the abilities of BS-align on 10 membrane proteins, whose lengths ranged between 229 and 595 residues, and whose high-resolution structures were difficult to determine, both by experiment and prediction. Our results confirmed that the backbone string neighbors generated by the BS-align algorithm were very close to the native membrane structures. Moreover, these conformational constraints dictated by the restricted regions of dihedral angles can be employed to guide the structure construction procedures such as in I-TASSER (16), Swiss-model (17), and Foldit online game (18).

MATERIALS AND METHODS

**Backbone String Prediction**—The flowchart of the backbone string prediction is shown in [supplemental Fig. S3](#). The PSI-BLAST algorithm was initially employed to match a query sequence against a protein database constructed by a nonredundant PDB chain set nr3PDB, resulting in two parts: matched fragments and unmatched fragments. Then, we utilized the hallmark patterns in the hallmark pattern library (see next) to hit the unmatched fragments and obtain the hit segments. These hit segments and their flaking amino acid residues (\( +n \) and \( -n \), default is 5) were aligned together against nr3PDB using PHI-BLAST (19), which found more matched shorter sequences. The matched fragments obtained by the first alignment and the shorter sequences obtained by the subsequent alignments were encoded based on corresponding backbone string element profile (see [supplemental Fig. S5](#)). The backbone string element profile was composed of eight elements (\( S, R, U, V, K, A, T, \) and \( G \)), which were employed as features for predicting the backbone string of the query. Lastly, conditional random field (see next) was performed for modeling and prediction.

**Hallmark Pattern Generated**—One innovative character of our approach was a knowledge-driven sequence alignment guided by seeds in a constructed hallmark pattern library (HPL), which was instrumental in searching structural similarities among highly divergent proteins. A hallmark pattern is composed of short consecutive sequences that are conserved both in the sequences and backbone strings. The HPL was the kernel of our approach and believed to reflect remote homologue information in the “twilight zone.” There are three steps to construct a hallmark pattern library to infer remote structural similarity among proteins.

Initially, we began a traversal search for consecutive sequence patterns with sufficient frequency in a representative nonredundant PDB chain set (nr3PDB, NCBI MMDB 2009 Dec, 0-level nonredundancy, 7,775 entries in total). In our previous study (20), we introduced an algorithm that could extract local combinational variables with fixed locations from equal-length sequences. Here, the algorithm was developed to extract candidate patterns from unequal-length sequences without sequence alignment (see [supplemental Fig. S4](#)). These short patterns were merged with every other single fragment that contained the same residue as the former fragment in order to form potentially longer sequences while maintaining the frequency criterion. We set the frequency criterion to 100 and a total of 5,667 consecutive sequence patterns were obtained. The entire pattern extraction process progressed as the fragment grew longer, a process known as the bottom-up method.

Second, hallmark patterns were defined as conservative both in sequence patterns and backbone string structures. For each position of a consecutive sequence pattern, the \( p \) value of the corresponding backbone string of residue at this position was calculated according to a binomially distributed model (see Eq. 1 below).

\[
P(N, m, q) = \sum_{i=m}^{N} \binom{N}{i} q^i (1 - q)^{N-i}
\]

Where \( N \) denoted the occurrence number of the pattern in the nr0PDB, a representative nonredundant PDB chain set; \( m \) denoted the count of maximum occurrence backbone string at position \( j \); and \( q \) denoted the corresponding backbone string background probability of residue at position \( j \). If one of the \( p \) values was less than 10\(^{-16}\), the consecutive sequence pattern was identified as a significant hallmark pattern.

Third, based on the \( p \) values, we selected 2761 hallmark patterns that typically exhibited conserved structures to construct the library. The HPL represented remote homology in the sequences and backbone strings and was an indispensable procedure in our approach.

**Backbone String Element Profile**—Another innovative character of our approach was the concept of backbone string element profile that was used for encoding as features for prediction. The backbone string element profile of a query was generated as follows: In the first step, the query sequence was aligned against the nr3PDB resulting in the top \( n \) subjects. The backbone strings of the \( n \) subjects were retrieved from BSD. Finally, the backbone string elements for each residue were counted and stored in eight boxes. These boxes constituted a vector that represents the backbone string element profile for each residue and was considered to include the structural evolutionary information. The backbone string element profile was utilized as features for modeling and prediction (more details about one-dimensional structural profiles can be found in our previous study (21)).

**Conditional Random Fields**—CRFs are frameworks for building probabilistic models to segment and label sequence data, based on the conditional approach over label sequences given a particular observation sequence \( X_{\text{test}} \) and label a novel observation sequence \( X_{\text{test}} \) by selecting the label sequence \( Y_{\text{test}} \) that maximizes the conditional probability \( P(\hat{Y}_{\text{test}}|X_{\text{test}}) \). Here, the features were formed only
by the backbone string element profile and the label referred to the backbone string for each residue.

Modeling—We used the “alignment mode” in the Swiss-Model (17) workspace to model the target protein. The aligned query and the template sequences were submitted to Swiss-Model. For 10 backbone string neighbors, the best z-score model was adapted as the final model of the query.

Software Availability—Software for protein backbone string prediction and protein backbone string alignment can be found at http://code.google.com/p/bs-align/.

RESULTS

In the present study, the backbone string representing a protein structure constitutes discrete regions in Ramachandran space. In the literature, this was also referred to as a shape string (15) or a one-dimensional string (22). The backbone string distribution and abundance is presented in supplemental Fig. S1 and supplemental Table S1. Fig. 1 illustrates how this backbone string is assigned to the backbone conformation of a protein and its neighbor.

Performance of Backbone String Prediction—The training data set, containing 4234 chains, was derived from the PDB (24) released before 2010 and was determined by x-ray diffraction with a resolution of $\leq 2.0\,\text{Å}$, an R-factor of $\leq 0.25$ and was cutoff at 25% sequence identity using PISCES (25).

Table I lists the performances of a five-fold cross-validation on 4234 nonredundant chain sets. The first row in Table I is the Segment Overlap (SOV) (26) measure, which is one of the prediction evaluation criteria for critical assessment of techniques for protein structure prediction. To calculate overall accuracy for three-state backbone strings (S3), we mapped eight-state backbone strings (S8) to the three-state by $S, R, U, V, A, K, H$, and $T, G$. The proposed method achieved an overall per-residue accuracy for the three-state backbone strings and eight-state backbone strings of 88.7% and 80.9% respectively, and an SOV of 86.4%, which was very close to the theoretical upper limit of accuracy of the secondary structure prediction (27).

To assess our backbone string prediction (BSP) method and the effect of the hallmark patterns, we used the latest EVA set (28) as an independent test set, which contained 79 proteins (1 was abolished out of 80 entries in the EVA set). The detailed results are listed in supplemental Table S2.

Fig. 1. Backbone string representation of protein structure. The corresponding backbone strings [S (red), R (green), U (blue), V (yellow), K (magenta), A (cyan), T (orange), G (wheat)] of amino acid residues are colored on both proteins. N denotes missing backbone string where the phi or psi angle is undefined, or gap in a PDB entry where no structure determination for parts of the sequence. A, The target protein (PDB: 1BYFA) and a backbone string neighbor (PDB: 3HR4A) obtained by aligning the predicted backbone string of 1BYFA against a backbone string database (BSD), which shared 23.4% sequence identity. Some residues (colored gray) of 1BYFA, which were not located in the x-ray diffraction experiment, share a certain degree of backbone string similarity with 3HR4A. Images of (A) and (B) were produced using PyMOL software (23). Seq: sequence, BS: backbone string, Pred_BS: predicted backbone string.

| Backbone string | S | R | U | V | K | A | T | G | Total |
|-----------------|---|---|---|---|---|---|---|---|-------|
| SOV (%)         | 84.5 | 86.9 |
| S3 (%)          | 79.8 | 69.8 | 53.1 | 48.5 |
| S8 (%)          | 89.2 | 91.5 | 57.5 | 91.4 | 76.5 | 77.1 | 65.7 | 80.9 |
duction by our method produced superior SOV (82.0%) and S3 (83.6%) values, outperforming an existing state-of-the-art method, Frag1D (22), by at least 6.9% in S3 and 4.7% in SOV. The more difficult S8 measure showed a remarkable improvement in performance (S8 74.4%, outperforming Frag1D by 6.8%) as well. The same trend occurred when the hallmark patterns were employed (outperforming when the hallmark patterns are not used by 9.2% in S3 and 6.2% in S8).

To assess the BSP method on newly measured proteins, we constructed independent test data by retrieving protein data released in the year 2010 from PDB, which were determined by x-ray diffraction with a resolution of ≤ 2.0 Å, an R-factor of ≤ 0.25, culled at 25% sequence identity, and contained 916 chains. Our method achieved an S3 of 84.6% and an S8 of 75.3%. The accuracy of prediction on three-state backbone strings S, H and T were 82.0%, 88.4% and 69.5%, respectively (see supplemental Table S3). The performance on newly measured PDB data demonstrated that the proposed method can be used for accurate prediction of protein backbone strings.

**Backbone String Database**—We utilized the actual backbone string of all known structural proteins in the nr3PDB database (NCBI MMDB 2009 Dec, three-level nonredundancy, 40849 entries in total) and constructed the backbone string database (BSD), which served as a benchmark alignment database. When we reduced the redundancy of the BSD by CD-HIT (29), the number of entries decreased quickly (Fig. 2), which confirmed the fact that the backbone string was more conserved than the sequence. These observations indicated that the backbone string maintained strong structural integrity and could be considered as the bridge between sequence-based and structure-based methods. When the backbone string identity was reduced to 50%, the number of left entries was approximately equal to the number of the folds in SCOP (1193, V1.75, 2009) (30). This finding implied that the backbone string may be a good criterion of protein classification. Moreover, the similarity of the backbone strings was the foundation of BS-align and was especially useful when sequences alignment was unfeasible.

**Performance of BS-align**—A nonredundant set of 620 sequences (31) was employed to test the performances of threading and modeling approaches. Fig. 3 presents the results of different methods on this benchmark set, where all homologous templates with sequence identity to targets >30% have been removed from the template library (BSD). The average TM-score (12) (TM-scores on the order of 0.5 are indicative of highly significant structural similarity), average root mean square deviation (RMSD), and average alignment coverage of all 620 queries were used to evaluate the performance.

The first model of BS-align achieved an average TM-score of 0.5671 and RMSD of 2.93 Å. Both were greatly improved in comparison to the best state-of-the-art method (TM-score = 0.4287 and RMSD = 6.92 Å). The average coverage achieved by the proposed method was 0.841.
Retrieving Membrane Protein Neighbors by BS-align—To demonstrate the ability of BS-align to retrieve candidates of backbone string neighbors from the sequences and use these candidates to model near-native structures of membrane proteins, we explored 10 large (>200 residues) and topologically complex membrane proteins from newly published references as a blind test set. Table II lists the results of the predicted neighbors of the best z-score (39) model for each membrane protein. The performance of the models was measured by RMSD of Cα between the model and the experimentally determined structure and the TM-score. In all cases, the models had RMSDs lower than 5.1 Å and TM-scores greater than 0.68. The coverage was more than 81.8% and the sequence identity was less than 40%.

CBB3 Cytochrome Oxidase—The structure of a C-family heme-copper oxidase (HCO) (PDB ID 3MK7) was determined by applying favorable anomalous properties of iron for phase determination at a resolution of 3.2 Å (40). The prediction was extremely accurate with an RMSD of 3.38 Å and a TM-score of 0.80 (Fig. 4A). The length of 3MK7 was 474 residues and...
Identifying Structural Neighbors from Primary Sequence

shared a 16.0% pairwise sequence identity with the neighbor protein recombinant cytochrome bα Oxidase (PDB ID 1XME). Whereas highly divergent in the amino acid sequence, different types of HCOs shared a similar structure, suggesting a similar mechanism of action (40, 41).

PorB—Neisseria meningitidis PorB (PDB ID 3A2R) was determined at 2.3 Å by using isomorphic and anomalous contributions for resolution of three heavy atom derivatives (42). The prediction was relatively good with an RMSD of 3.78 Å (Fig. 4D) and a TM-score of 0.68, which only shared a 16.6% pairwise sequence identity with the backbone string neighbor, protein Escherichia coli K-12 (PDB ID 3HWB). PorB was previously predicted to share a stable, 16-stranded β-barrel scaffold with other porins; however, even superimposition of PorB with known structures resulted in a low RMSD value, only 20 absolutely conserved residues were identified (42), which made primary sequence-based structure neighbor identification extremely difficult.

Carnitine transporter protein (CaiT)—The structures of the sodium independent carnitine/butyrobetaine antiporter CaiT from Proteus mirabilis (PmCaiT) (PDB ID 2WSW) was previously solved at 2.3 Å resolution by molecular replacement with a poly-alanine model of BetP (PDB ID 2WIT) (43). Here, we showed it was possible to find the more native neighbor and predicted PmCaiT at a high resolution with RMSD of 2.08 Å (Fig. 4B).

ATP-binding cassette transporter—Prediction of the first structurally-determined human ATP-binding cassette transporter (PDB ID 2YL4) was relatively less well-defined with an RMSD of 5.03 Å and a TM score of 0.73 (Fig. 4C). This result may have been due to the large size of the protein (595 residues) and a high level of topological complexity, which consisted of six transmembrane domains and two cytoplasmic nucleotide-binding domains. The other six targets are illustrated in supplemental Fig. S2.

DISCUSSION

Hallmark Pattern—Despite the developments in discovering motifs, little research has focused on the relationship between sequence patterns and their corresponding structures. The hallmark pattern that we propose in this study is a union of sequences and backbone strings. We used the short, conformational restricted consecutive sequences as a seed to guide sequence alignment in unmatched sequences, which were typically considered as remote homolog regions. It was our insight that when the sequence similarity was low, the knowledge-driven method produced better sequence alignments than using sequence similarities alone. The HPL (see Methods) is a kernel library of the BS-align algorithm and we believe that the HPL will be beneficial in finding remote homology in the “twilight zone” (≤25% similarity).

The Advantages of the Backbone String—There are three advantages of the backbone string that was introduced into the field of machine learning. One of the most prominent advantages of backbone string was its ability to describe the detailed protein backbone structure. Many studies have taken advantage of backbone conformational information (44–46). Our previous work (47, 48) demonstrated that backbone string was important for turn identification as well. The second advantage was that the backbone string was more conservative than the sequence. For BSD, when the backbone string identity was reduced to 40% (Fig. 2), only 74 entries (72 proteins) remained, which indicated that cross fold similarities were abundant in geometrically similar proteins. Based on the SCOP classification system, there were 24 all-beta proteins, 14 alpha and beta proteins (a+b), 14 small proteins, 12 alpha and beta proteins (a/b), 10 all-alpha proteins, four multidomain proteins, three membrane and cell surface proteins, four peptides and one coiled coil protein found in these entries with lengths varying between 54 and 2512 residues. This phenomenon implied that the protein structures were fairly conserved and suggested that the backbone string may be a suitable criterion of taxonomy and a backbone string-based library may be more reasonable and compact than existing libraries. The third advantage was the alignment manipulation of the backbone string, which was relatively simple to complete, due to the backbone string being composed of eight elements rather than accurate backbone torsion angle values. The alignment accuracy between a query protein sequence and a known template structure was the key in determining the accuracy of the final three-dimensional model and was a bothersome procedure. The sequence alignment may not be of sufficient structural quality, whereas the secondary structure element alignment (SSEA) was inappropriate for accurate structural alignment. The backbone string alignment, described herein, demonstrated its ability to align against a database using sophisticated algorithms, such as BLAST, to produce accurate results.

Backbone String Neighbors—The backbone string neighbor is a structural neighbor based on sequence, and represents a new bridge joining sequence to structure. Despite the crucial functions performed by membrane proteins in living cells, it remains frustratingly difficult to obtain high-resolution 3D structures of membrane proteins (49). We identified the backbone string neighbors of membrane proteins and showed that our algorithm could build models of membrane structures from sequences even if they have large sizes and complex topologies. Because structural and functional characterization of membrane proteins relied on the isolation and purification of large amounts of sequences, we believe that BS-align will play an important role in computational biology both in prediction approaches and combining experimental approaches.

* This work was supported by the National Natural Science Foundation of China grants (20675057, 20705024).
≤ This article contains supplemental Figs. S1 to S5 and Tables S1 to S3.
§ To whom correspondence should be addressed: Department of Chemistry, Tongji University, 1239 Siping Road, Shanghai 200092, China. Tel.: 86 21 65983987; E-mail: lilith@tongji.edu.cn.

10.1074/mcp.M111.016808–6 Molecular & Cellular Proteomics 11.7
In order to cite this article properly, please include all of the following information: Sun, J.-M., Li, T.-H., Cong, P.-S., Tang, S.-N., and Xiong, W.-W. (2012) Retrieving Backbone String Neighbors Provides Insights Into Structural Modeling of Membrane Proteins. *Mol. Cell. Proteomics* 11(7):M111.016808. DOI: 10.1074/mcp.M111.016808.