Tertiary Structure Prediction of the KIX Domain of CBP Using Monte Carlo Simulations Driven by Restraints Derived From Multiple Sequence Alignments

Angel R. Ortiz,[1] Andrzej Kolinski,[1,2] and Jeffrey Skolnick[1]*
[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*
[2]*Departmentof Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **Using a recently developed protein folding algorithm, a prediction of the tertiary structure of the KIX domain of the CREB binding protein is described. The method incorporates predicted secondary and tertiary restraints derived from multiple sequence alignments in a reduced protein model whose conformational space is explored by Monte Carlo dynamics. Secondary structure restraints are provided by the PHD secondary structure prediction algorithm that was modified for the presence of predicted U-turns, i.e., regions where the chain reverses global direction. Tertiary restraints are obtained via a two-step process: First, *seed* side-chain contacts are identified from a correlated mutation analysis, and then, a threading-based algorithm *expands* the number of these seed contacts. Blind predictions indicate that the KIX domain is a putative three-helix bundle, although the chirality of the bundle could not be uniquely determined. The expected root-mean-square deviation for the correct chirality of the KIX domain is between 5.0 and 6.2 Å. This is to be compared with the estimate of 12.9 Å that would be expected by a random prediction, using the model of F. Cohen and M. Sternberg (J. Mol. Biol. 138:321–333, 1980). Proteins 30:287–294, 1998.** © **1998 Wiley-Liss, Inc.**

Key words: protein structure prediction; side chain contact prediction; lattice protein models; CREB-binding protein; KIX domain

INTRODUCTION

Recently, encouraging progress has been made in the ab initio folding of small proteins by using predicted secondary and tertiary restraints to assemble global protein topologies[1] (Ortiz, Kolinski, and Skolnick, manuscripts submitted). In this approach, secondary structure prediction schemes provide the information required to describe the local chain conformation.[3–5] Tertiary contacts are predicted on the basis of evolutionary information contained in multiple sequence alignments.[5] These are supplemented by additional contacts extracted from a threading procedure.[6–7] The predicted set of restraints is then included as a soft biasing potential in lattice Monte Carlo simulations that also incorporate statistical potentials to drive the assembly process.[7] This combined approach seems to be able to bridge the gap between sequence analysis and folding simulations and permits the ab initio folding of some rather complex topologies.[2] Although encouraging results have been obtained using a representative test set of 20 different small proteins, validation of this approach requires bona fide blind predictions of protein structures that will be subsequently obtained by independent experiments. In this spirit, here we describe the prediction of the KIX domain of the CREB-binding protein. This molecule is involved in gene expression mediated by cAMP. Its structure has just been solved by nuclear magnetic resonance in Dr. Peter Wright's group at The Scripps Research Institute, but he has not provided us with any information about what this structure might be.†

It has been firmly established that hormonally induced increases in cAMP levels stimulate gene expression. This cAMP-regulated gene expression frequently involves a DNA element known as the cAMP-regulated enhancer (CRE). CRE is an octanucleotide motif (TGACGTCA) that mediates diverse transcriptional regulatory effects. Many transcription factors bind to this element, including CREB, which is activated as a result of phosphorylation by protein kinase A. Upon phosphorylation,

†A letter from Dr. P. E. Wright has been transmitted to the editor stating that he has not provided us with any structural information about the KIX domain (1997).

Contract grant sponsor: National Institutes of Health, Contract grant number: GM-37408; Contract grant sponsor: University of Warsaw, Contract grant number: BST-34/97 [A.K.]; Contract grant sponsor: Spanish Ministry of Education [A.R.O.]

*Correspondence to: Dr. Jeffrey Skolnick, Department of Molecular Biology, TPC-5, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037.

Received 1 July 1997; Accepted 22 September 1997

CREB binds as a dimer to CRE and activates gene transcription. Once activated, CREB also specifically binds to a nuclear protein of M(r) 265K termed CBP (for CREB-binding protein).[9] In particular, it binds to a domain of CBP known as KIX. Thus, CBP may participate in cAMP-regulated gene expression by interacting with the activated phosphorylated form of CREB. CREB transcription factors are about 350 amino acid residues long. They contain a COOH-terminal leucine zipper motif and an adjacent basic domain responsible for transcriptional activity. Regulation by phosphorylation of CREB takes place at a serine residue in the phosphorylation domain. This 60-amino-acid region is termed the kinase-inducible domain (KID)2.[10] It has been shown that the KID domain interacts with the KIX domain of CBP. KIX is a domain of 78 amino acids and, therefore, it is within the size range amenable to ab initio folding. In the rest of this article, we provide a detailed description of the tertiary structure prediction of the KIX domain.

METHODS

Overview

A flow chart of the tertiary structure prediction protocol is schematically depicted in Figure 1. The procedure can be logically divided into two parts: restraint derivation and structure assembly/refinement using our recently developed MONSSTER algorithm.[8] With respect to restraint derivation, the first objective is to predict the number, location, and identity of the dominant secondary structural elements that will comprise the folded protein. These helices and β strands comprise the core topological elements of the molecule. In addition, U-turns between these core secondary structure elements are predicted.[5] Next, we try to predict which secondary structure elements are in contact. First, we obtain the most reliable set of predicted contacts between core elements from correlated mutation analysis.[11] We denote such contacts as *seeds*. Next, we exploit the fact that packing patterns between secondary structure elements are degenerate. Given the predicted seed and putative secondary structure partners, seed contacts are *enriched* by an inverse folding/fragment clustering protocol.[6] Then, an updated version of MONSSTER designed to accommodate the inherent inaccuracies of such restraints assembles the fold.[2]

Restraint Derivation

Multiple sequence alignment

A multiple sequence alignment was obtained for the KIX sequence by scanning the EMBL/SWISS-PROT database with FASTA[12] and filtering the sequences found using MAXHOM.[13] After further filtering the alignment by hand, the final multiple sequence alignment contained 58 homologous sequences plus the target sequence (Fig. 2).

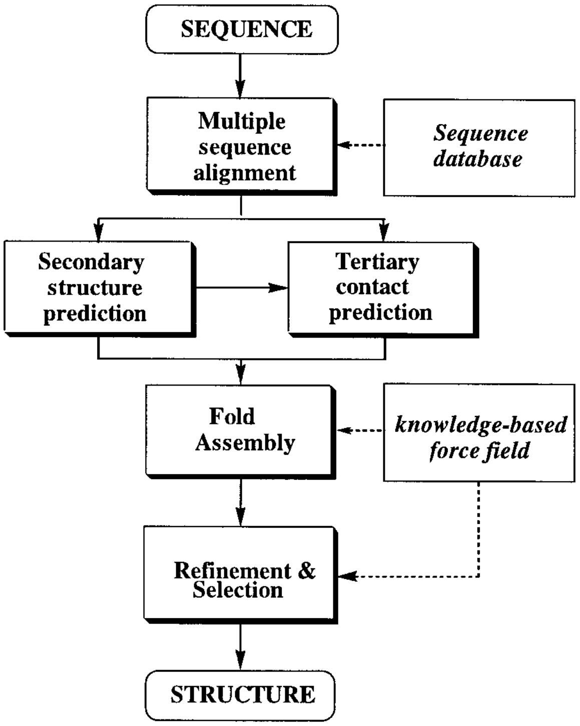
SEQUENCE
Multiple sequence alignment
Sequence database
Secondary structure prediction
Tertiary contact prediction
Fold Assembly
knowledge-based force field
Refinement & Selection
STRUCTURE

Fig. 1. Flow chart of the protein fold prediction method.

Secondary structure prediction

The multiple sequence alignment given in Figure 2 was used as input for the PHD secondary structure prediction method. For helices, only those elements with a reliability index higher than 3 are used. Then, chain reversals are predicted by the *U-turn* prediction algorithm LINKER developed by Kolinski and coworkers.[5] Because of their reliability, elements predicted as U-turns override PHD predictions.[5] Thus, each residue is assigned to be in one of five conformational states: a predicted extended/loop state, a predicted helix, a predicted U-turn, a β (strand) state or a nonpredicted state. The set of predicted helices and strands comprise the putative core elements of the protein. Figure 3 presents the resulting secondary structure prediction.

Side-chain contact prediction

Residue contact prediction is performed in two stages: First, a correlated mutation analysis[11] of the multiple sequence alignment is done to identify the *seed* contacts. In the calculation of the covariance matrix, regions containing deletions and insertions are not considered. Residue comparison is carried out using the McLachlan matrix.[14] Only correlations between predicted core elements are considered, and correlation is measured by a Pearson-type correla-

```
NR.    ID                                      PROTEIN_SEQUENCE
 0 : cbp_mouse     GVRKGWHEHVTQDLRSHLVHKLVQAIFPTPDPAALKDRRMENLVAYAKKVEGDMYESANSRDEYYHLLAEKIYKIQKELEE
 1 : ynj1_caeel       KEWHHQVTKDLRNHLVGKLVKAIFPEPNQEAMNDNRLKDLIAYARKVEKEMFESANDREEYYHLLAEKIYKIQKELQE
 2 : myse_chick       KKNLDQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEgyERRVKELTYQSEEDRKNVLRLQdsKFRKIQHELEE
 3 : myse_human       KKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLETRIRELEfyERRVKELTYQSEEDRKNVLRLQdtKFRKAQHELEE
 4 : myse_rat         KKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLETRIRELEfyERRVKELTYQSEEDRKNVLRLQdtKFRKAQHELEE
 5 : mysp_human       KKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEghDRRVKELTYQTEEDRKNVLRLQdsKFRKLQHELEE
 6 : myss_chick       KKNMDQTVKDLQLRLDEAEQLALKGGKKQLQKLEARVRELEgyERRVKELTYQCEEDRKNILRLQdsKFRKIQHELEE
 7 : mysp_rat         KKNMEQTVKDLQHRLDEAEQLALKGGKKQIQKLEARVRELEghERRVKELTYQTEEDRKNVLRLQdaKFRKLQHELEE
 8 : myss_rabit       KKNMEQTVKDLQQRLDEAEQLALKGGKKQIQKLEARVKELEnhERRVKELTYQTEEDRKNVLRLQdsKFRKLQHELEE
 9 : myss_human       KKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLEAKVRELEghERKVKELTYQTEEDRKNILRLQdsKFRsIQHELEE
10 : myse_mouse       KKNLEQTVKDLQHRLDEAEQLALKGGKKQIQKLETRIRELEfyERRVKELTYQSEEARKNVLRLQdtKFRKAQHELE
11 : mysb_mesau       KKNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEnsERRIKELTYQTEEDRKNLLRLQdsKFRKVQHELDE
12 : mysb_human       KKNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEnsERRIKELTYQTEEDRKNLLRLQdsKFRKVQHELDE
13 : mysb_papha       KKNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEnsERRIKELTYQTEEDRKNLLRLQdsKFRKVQHELDE
14 : mysa_human       KKNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEgsERRIKELTYQTEEDKKNLLRLQdsKFRKVQHELDE
15 : mysb_rat         KNNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEnsERRIKELTYQTEEDRKNLLRLQdsKFRKVQHELDE
16 : mysa_rat         KKNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEnsERRIKELTYQTEEDKKNLVRLQdsKFRKVQHELDE
17 : mysa_mesau       KKNMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELEnsERRIKELTYQTEEDKKNLVRLQdsKFRKVQHELD
18 : mysc_chick       KKNMEQTIKDLQKRLDEAEQIALKGGKKQIQKLESRVRELEnfERRIKELTYQSEEDKKNLARMQdsKYRKQQHDLDD
19 : mysa_drome       RKALEQQIKELQVRLDEAEANALKGGKKAIQKLEQRVRELEnsERRVKELSFQSEEDRKNHERMQdaKFRKAQQELEE
20 : mys_aeqir        RKNLESQVKEFQIRLDEAEASSLKGGKKMIQKLESRVHELEaaDRRLKELAFQADEDRKNQERLQeaKYRKAQHELEE
21 : mysb_caeel       RKGLEQQLKEIQVRLDEAEAAALKGGKKVIAKLEQRVRELEsaDRRVRELQFQVDEDKKNFERLQdqKLKTQKKQVEE
22 : mysp_drome       KKSLEVEVKNLSIRLEEVELNAVAGSKRIISKLEARIRDLElkERTVKEVLVQCEEDQKNLILLQdtRVRRFQRELEA
23 : mysq_drome       KKSLEVEVKNLSIRLEEVELNAVAGSKRIISKLEARIRDLElkERTVKEVLVQCEEDQKNLILLQdtRVRRFQRELEA
24 : mysa_caeel       RKGLELQIKEMQIRLDDAENAALKGGKKIIAQLEARIRAIEqaERRVKEVEFQVVEEKKNEERLTecKLKIFKRQVEE
25 : mysp_schma       RKQLEIEIREITVKLEEAEASATREGRRMVQKLQARVRELEsfERQYKELQTQAEDDRRMVLELQdnKYRKAQQQIEE
26 : mysp_echgr       RKQLEIEIREITVKLEEAEAFATREGRRMVQKLQNRVRELEayERQFKELQTQSEDDKRMILELQdsKYRKAQQQIEE
27 : mysp_taeso       RKQLEIEIREITVKLEEAEAFATREGRRMVQKLQNRVRELEayERQFKELQTQSEDDKRMILELQdsKYRKAQQQIEE
28 : mysd_caeel       RKSLELNAKELQAKIDDAERAMIQFGAKALAKVEDRVRSLEaqERRARELQFQVEEDKKAFDRLQesKFRQIQLALEN
29 : mysp_caeel       RKSLEEQVKQLQVQIQEAEAAALLGGKRVIAKLETRIRDLEtkDRRIKEVQQLVDEEHKNFVMAQdqRVRRYQHELED
30 : mysp_dirim       RKSLEEQVKQLQVQIQEAEAAALLGGKRVIAKLETRIRDLEtkDRRIKEVQMQVDEEHKMFVMAQdqRVRRYQRELED
31 : mysp_oncvo       RKSLEEQVKQLQVQIQEAEAAALLGGKRVIAKLETRIRDLEtkDRRIKEVQMQVDEEHKMFVMAQdqRVRRYQRELED
32 : mysp_bruma       RKSLEEQVKQLQVQIQEAETAALLGGKRVIAKLETRIRDLEtkDRRIKEVQMQVDEEHKMFVMAQdqRVRRYQRELED
33 : mysc_caeel       KKQLESAVKDLQERADAAEAAVMKGGAKAIQKAEQRlkTLARADRKVREFEFQVAEDKKNYDKLQeaKLKLQKKQLEE
34 : mys_podca        RQLVSKQVADLQSRLEDAEAQGGKGLKNQLRKLEQRIMELEssEKKVKELAFTIEDEHKRREPAQdqKLKKMRMQLEE
35 : m5_strpy         KKILDETVKDKLakKDEANKISDASRKGLRRDLDAskKQLEAEHQKLEEQNKISEASRKGLRrlDASREAKKQLEAEQ
36 : mysb_rabit       KIQLEAKVKEMNERLEDEEEMNAeeLKRDIDDLELTLAKVEktENKVKNLtlQAEEDKVNTLTKAKVKLEQQVDDLE
37 : mysn_acaca       AKTLKTQVDETKRRLEEAEASAArsGAQQRRKLNTRISELQssSEEVKRLEGELERLEEELLTAQekNLDKANLELEE
38 : tpmx_rat         KKKMQMLKLDKENALDRAEqkAAEDRSKQlkKLKATEDELDKYSEALKDAQEKLELAEKKATDAEasLNRRIQLVEEE
39 : tpma_brare       KKKMQMLKLDKENALDRAEqkAAEERSKQlkKLKATEDELDKYSEALKDAQEKLELAEKKATDAEgsLNRRIQLVEEE
40 : mys2_dicdi       KRELEIRVEDMESELDEKK.LALENLQNQKRSVEEKVRDLEelRNTLEKLKKKYEEELEEMKRVNdsRLEKIKDELQK
41 : rpsd_myxxa       KKEVKQEIKDLRTKMMEVLeiNLKALIERVDKAEDELRDLEryDCSMKELRPQLKESRENP.....NIGKKLQKQLN

42 : tpmm_trico       QKKMMQTENDLDKAQEdaATSQLEEKEKKVQEAEAEveELERAEERLKIATEKLEEATHNVdfQDEERANTIEAQLKE
43 : eg5_human        KNELDQCKSDLQNKTQELeqKHLQETKLQLVKEEYITSALESTEEKLHDAASkvEETTKDVSGLHSKLDRKKAVD
44 : mys1_yeast       SEQLDRLQKDLE.STERQKELLSSTIKQQKQQFENCMDDLQGNELRLREHIhqAEEDVKNmlKTQNKQKEKLIWERE
45 : mysa_rabit       KLKLEQQVDDLEGSLEQEKKvdLKLTQESIMDLENDKQQLEekEFDISQLNSKIEDEQALVLQLQ.KKLKENQARIEE
46 : us45_lacla        ATLNESIKERTKTLEAQARSAqkSLTDVIQKVTanKQILEQQEKEQKELSQKSETVKKNylDSQAQELTSQQAELK
47 : ca36_chick       GKALDYVVKNYfqSQDDVNRPANVISSTSIQPLGVGARNVDRNQLQVIT......NDPGRVLVVQdtLERKVQNILEE
48 : ynj1_caeel       KEWHHQVTKDLRNHLVGKLVKAIFPEPNQEAMNDNRLKDLIAYARKVEKEMFESANDREEYYHLLAEKIYKIQKELQE
49 : tpm2_yeast       NEQLDSEVEKLESQLSDTKQLatKKNQDLEQQLEDSEAKLKEAMDKLKEADLNSEQMGRRIVALEesKYEEAQKELDE
50 : tpm2_human       KKKMQMLKLDKENAIDRAEQAEADKKQAeqKKLKGTEDEVEKYSESVKEAQEKLEQAEKKATDAEasLNRRIQLVEEE
51 : lama_chick       KRNLENEVRDLRAQVAKLEG.ALSEAKKQLqkRRHETRLVEIDNGRQQEFESKLAEALQDLRRQHEDQIRHYRDELEK
52 : tpmb_chick       KKKMQMLKLDKENAIDRAEQAEADKKQAeqKKLKGTEDEVEKYSESVKEAQEKLEQAEKKATDAEasLNRRIQLVEEE
53 : tpms_cotja       KKKMQMLKLDKENALDRAEQAEakQLEDDIVQLEKQLRVTEDSRDQVLEELHKSEDSLLSAEEIAasLNRRIQLVEEE
54 : tpmb_human       KKKMQMLKLDKENAIDRAEQAEADKKQAeqKKLKGTEDEVEKYSESVKEAQEKLEQAEKKATDAEasLNRRIQLVEEE
55 : tpm1_chick       KKKMQMLKLDKENAIDRAEQAEADKKQAeqKKLKGTEDEVEKYSESVKEAQEKLEQAEKKATDAEasLNRRIQLVEEE
56 : rpoa_bacsu       SKILTEHLNIFVGLTDEAQHAEIMVEKEEDQklEMTIEELDLSVrtVQELANKTEEDMMKVRNLGRKSLEEVKAKLEE
57 : invo_ponpy        KHLDQQEGQLKHLDQQEKQLELPEQQvqLKHLEQQEGQLEVPEEQVGQLKYLEQQEGQ....LKHLDQQEGQLELPE
58 : hs7c_caeel            LSPEDIEAMINDAEKFA.EDDKKVKDKAEAR.NELESYAYNLKN...QIEDKEKLGGKLDEDDKKTIEEAVEE
```

Fig. 2. Multiple sequence alignment used in the secondary structure and tertiary contact predictions. NR, sequence number in the alignment; 0, target sequence; ID, identification number according to the EMBL/SWISSPROT database (except the target sequence); PROTEIN_SEQUENCE, the sequence alignment used for contact prediction; lower case letters indicate that in that position an insertion (not shown) is found in the corresponding sequence.

tion coefficient.[11] A correlation coefficient cutoff threshold of 0.5 is used for contact prediction.

In general, only one contact per each pair of core secondary structure elements is used. However, in some cases, use of more than one contact per pair of secondary structure elements may be advantageous if the elements of the set of seed restraints are consistent with each other, particularly when there is a strong asymmetry in the final restraint distribution, as in the present case (see below). The set of seed restraints is then enriched by a combined threading and structural fragment procedure.[6] All pairs of secondary structure elements compatible with the predicted secondary structure types and contact locations are extracted from a structural database. To account for the inaccuracies in the correlated mutation analysis, a ±1-residue shift in each member of the contacting pair is allowed. Fragments are then scored by a potential that considers local conformational propensities and the burial energy within the pair of fragments.[6,7] The top 10 scoring fragments are superimposed in space. If they

```
          10        20        30        40        50        60        70        80
          |         |         |         |         |         |         |         |
|GVRKGWHEHVTQDLRSHLVHKLVQAIFPTPDPAALKDRRMENLVAYAKKVEGDMYESANSRDEYYHLLAEKIYKIQKELEE|
|-----HHHHHHHHHHHHHHHHHHHHH-----HHHHHHHHHHHHHHHHHHHHHHHHH----HHHHHHHHHHHHHHHHHHH--| PHD
|-----------------------------UU--------------------UU----------------------------| LINKER
|-----HHHHHHHHHHHHHHHHHHHHH---UUHHHHHHHHHHHHHHHHHHHHUUHHH----HHHHHHHHHHHHHHHHHHH--| PHD+LINKER
|               I                        II                          III          |
```

Fig. 3. Secondary structure prediction of the KIX domain. The protein sequence is shown, together with the predicted secondary structure, by using the following algorithms: First, the results of the PHD prediction are shown, with an H indicating those positions where helix prediction has a reliability index >3. Second, the results of the U-turn prediction, i.e., those regions in which the chain reverses direction, are shown by using the LINKER program. In the next line, the PHD and LINKER predictions are combined in a unique prediction that was used in the actual folding simulations (see the Methods section for additional details).

do not show a clear spatial clustering (with an upper limit of 5.5 Å for the most divergent fragment pair), then enriched side-chain contact restraints are not derived. Conversely, if the fragments spatially cluster, then the fragment within this cluster whose root-mean-square deviation (RMSD) is smallest, with respect to all other members, is selected, and its side-chain contact map is projected onto the query sequence.

Structural Assembly

Protein representation

The Cα coordinates of the protein backbone are confined to a set of lattice points located on an underlying cubic lattice whose lattice spacing a = 1.22 Å.[15] Successive Cα atoms are connected by a set of 90 virtual bond vectors $\mathbf{a} \cdot \mathbf{v}$, with $\{\mathbf{v}\} = \{(\pm 3, \pm 1, \pm 1), \ldots (\pm 3, \pm 1, 0), \ldots (\pm 3, 0, 0), \ldots (\pm 2, \pm 2, \pm 1), \ldots (\pm 2, \pm 2, 0), \ldots\}$. The distance a is chosen so that the mean Cα virtual bond length is 3.8 Å. Side chains are represented by a set of rotamers, each located at the side-chain center of mass, and are not restricted to lattice points. With the exception of Gly, Pro, and Ala, there are multiple rotamers for each amino acid.

Interaction scheme

There are two types of interactions in the model: inherent contributions and restraint contributions.[2] The first class of terms is independent of the restraint predictions and is designed to capture both generic (sequence-independent) and sequence specific proteinlike features.[8] Such terms include an amino acid pair-specific potential that describes the intrinsic secondary structural preferences and a one-body centrosymmetric burial potential. Also, in order to avoid nonphysical segregation of the subunits, we have added a packing density regularizer.[16] The side-chain pair contact potential has been derived by a careful analysis of the appropriate reference state.[17] Hydrogen bonds are Cα-based and are very much in the spirit of Levitt and Greer.[18]

Restraint contributions are based on the predicted secondary structures and tertiary contacts. There is a local bias for the predicted secondary structure type as well as hydrogen bond mixing rules that specify the type(s) of secondary structure that can form hydrogen bonds.[2] To account for the level of precision of the predicted contacts, the tertiary contact restraint function consists of a simple flat-bottom harmonic potential operating on the projection of the residue pair onto the principal axes of their respective secondary structural elements. The target distance is the average separation observed in a structural database for the packing of secondary structure elements. These values are 10 Å for contacting α-α, 8 Å for α-β and 6 Å for β-β supersecondary elements, respectively.

Conformational sampling

Sampling of conformational space occurs via a standard asymmetric Monte Carlo Metropolis scheme.[19] Several types of local conformational micro-modifications of the chain backbone are used.[2] In this work, 10 independent assembly simulations were carried out, each from a fully extended initial conformation. Each simulation starts at a reduced temperature of 5.0, and then, the temperature is slowly lowered to 1.0 (when using a statistical potential of mean force, energies are usually expressed in k_BT units. The simulation temperature is referred to in this system of units. Thus, a *reduced temperature* of 1.0 indicates that the simulation temperature corresponds to the fictitious temperature of the database used to derive the energy scale, whatever that temperature is). Low-energy structures are then subject to isothermal refinement. The predicted fold is the one exhibiting the lowest average energy during the isothermal calculation.

RESULTS AND DISCUSSION

Derivation of Restraints

Secondary structure restraints

A multiple sequence alignment was built comprised of 58 protein sequences plus the target sequence, whose identities ranged from 68% to 31% (data not shown; Figure 2 shows the actual multiple sequence alignment). This range of sequence identity has been found to be suitable for contact prediction.[20] Most of the detected sequences are related to myosin or other muscle proteins. Prediction of PHD and LINKER are in reasonable agreement with each other (Fig. 3), although the second U-turn, predicted

TABLE I. Predicted Seeds for the KIX Domain Extracted From Correlated Mutation Analysis

Residue number		Residue name		
Partner A*	Partner B*	Partner A	Partner B	Correlation coefficient†
22	35	Leu	Leu	0.707
22	73	Leu	Tyr	0.673
35	73	Leu	Tyr	0.623
17	72	His	Lys	0.633

*Residues "A" and "B" refer to the first and second partner of the predicted contact, respectively.

†Correlation coefficient for the mutational behavior of the corresponding positions in the multiple sequence alignment.

TABLE II. List of Predicted Contacts Used in the Structure Prediction of the KIX Domain

Res. A (T0042)*	Res. B (T0042)†	Template structure‡	Res. A (template)§	Res. B (template)¶
18	22	1ezm	269	273
18	39	1ezm	269	290
21	35	1ezm	272	286
21	39	1ezm	272	290
22	26	1ezm	273	277
22	32	1ezm	273	283
22	35	1ezm	273	286
22	36	1ezm	273	287
22	39	1ezm	273	290
23	26	1ezm	274	277
25	30	1ezm	276	281
25	32	1ezm	276	283
25	35	1ezm	276	286
28	30	1ezm	279	281
31	33	1ezm	282	284
31	34	1ezm	282	285
34	37	1ezm	285	288
35	39	1ezm	286	290
37	40	1ezm	288	291
37	41	1ezm	288	292
39	42	1ezm	290	293
32	76	1cpc_A	48	92
35	76	1cpc_A	51	92
36	69	1cpc_A	52	85
36	72	1cpc_A	52	88
37	69	1cpc_A	53	85
39	72	1cpc_A	55	88
40	68	1cpc_A	56	84
40	69	1cpc_A	56	85
40	72	1cpc_A	56	88
43	72	1cpc_A	59	88
44	65	1cpc_A	60	81
44	68	1cpc_A	60	84
51	65	1cpc_A	67	81
22	35	SEED	—	—
22	73	SEED	—	—
35	73	SEED	—	—
17	72	SEED	—	—

*First residue in the predicted contact in the target sequence of unknown structure.

†Second residue in the predicted contact in the target sequence of unknown structure.

‡Template structure from which the predicted contact is extracted. The protein name corresponds to the PDB entry. The last four contacts are obtained from the correlated mutation analysis.

§Residue number in the template structure of the first partner of the predicted contact.

¶Residue number in the template structure of the second partner of the predicted contact.

in positions 52-53, overrides the C-terminal end of helix II, splitting it, and leaving a small helix between residues 54 and 56. After combining the PHD and LINKER predictions, as shown in Figure 3, the predicted secondary structure suggests the putative presence of three long helices in the protein. Additionally, one small helix, three residues in length, may be present between residues 54 and 56. The reliability index for the predicted helical regions is rather high (data not shown).

Contact prediction

Three *seed* contacts, shown in Table I, can be derived from the correlated mutation analysis.[11] These are then enriched to give the total predicted contact set shown in Table II. Enriched contacts could be extracted for two of the seed contacts, involving helices I and II and helices II and III (see helix definitions in Fig. 3). However, for helices I and III, it was not possible to derive a consensus packing pattern from the structural database. For this reason, and to partly avoid the imbalance in the restraint energy among the different regions of the structure, the next predicted contact between helices I and III was also selected and used in the folding simulations (Table I). Since the contact was compatible with the previous one, we felt reasonably confident about the contact prediction on this region. The resulting number of predicted contacts, 38, is rather high as compared to typical results on similar size proteins.[2]

Fold Assembly and Discrimination

All 10 runs readily converged to one of two different topologies, either a left- or right-handed, three-helix bundle. No other topologies are observed. The handedness of the bundle is defined as follows: Let the principal axis of helix I be oriented parallel to the *z* axis, with its C-terminal in the positive *z* direction. Then, let the principal axis of helix II lie in the *xz* plane. If helix III lies below this plane (i.e., has negative *y* values), then this constitutes a left-handed, three-helix bundle. Conversely, if helix III lies above this plane, then it is a right-handed, three-helix bundle. We term the left- and right-handed three-helix bundles topological mirror images, since these are folds where the chirality of the secondary structural elements is the same, but the chirality of the turns is reversed.[21]

At the end of the assembly runs, the residual restraint energy was essentially negligible for both

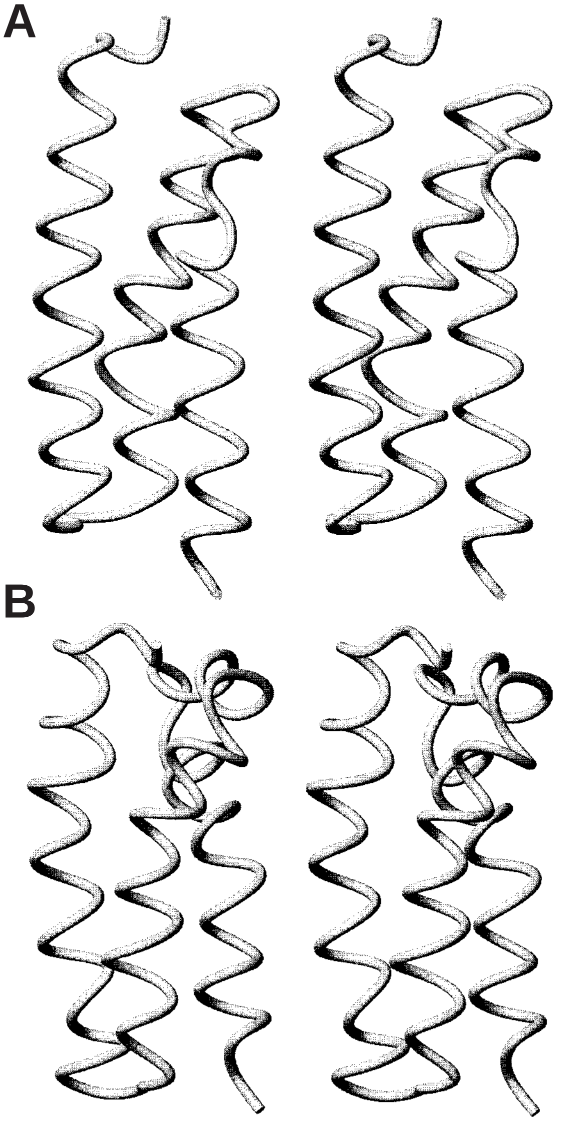

Fig. 4. Predicted structures of the KIX domain. **A:** Right-handed, three-helix bundle. **B:** Left-handed, three-helix bundle. Stereo figures generated with MOLMOL.[21]

TABLE III. Average Energies and Geometric Characteristics of the Two Topological Mirror Images of the Predicted Fold of the KIX Domain for the Lowest Average Energy Structural Representative of Each Topology*

Energy[†]	Left-handed bundle	Right-handed bundle
$\langle S^2 \rangle$	107.39	119.24
$\langle$%Helicity$\rangle$	75.80	75.80
rs/*rt*	19/37	26/37
$\langle E \rangle_{\mathrm{tot}}$	**−476.33**	**−478.56**
σ	6.90	7.12
$\langle E \rangle$	−495.63	−510.06
$\langle E \rangle_{\mathrm{res}}$	0.80	0.03
$\langle E \rangle_{\mathrm{b+r}}$	19.30	31.50

*All energies are in kT units.

[†]$\langle S^2 \rangle$ is the average squared radius of gyration computed during the isothermal run. The expected value, taking into account the number of residues and considering a spherical model, is 91.74.

$\langle$%helicity$\rangle$ is the average value of the percentage of residues in the helical state during the isothermal run.

rs/*rt* is the number of restraints satisfied in the final fold with respect to the number of target restraints used in the simulations.

$\langle E \rangle_{\mathrm{tot}}$ corresponds to the average total energy obtained from the isothermal run at $T = 1.0$.

σ is the standard deviation of the total energy from the isothermal run at $T = 1.0$.

$\langle E \rangle$ is the average of the total energy minus the contributions of the burial and density regularizing terms from the isothermal run at $T = 1.0$.

$\langle E \rangle_{\mathrm{res}}$ is the average restraint energy from the isothermal run at $T = 1.0$.

$\langle E \rangle_{\mathrm{b+r}}$ is the average value of the burial term together with the density regularizer component from the isothermal run at $T = 1.0$.

topological mirror images. Isothermal stability calculations were carried out in order to distinguish between the two possibilities; a ribbon representation of each of the predicted folds is shown in Figure 4. As shown in Table III, the energy differences between the two topologies did not allow us to reject either solution. There are three different characteristics of the protein that could, in principle, be different in simplified representations of topological mirror images. One is the burial pattern, the second is the pair interactions and the third is related to the secondary structure propensities for the different chiralities of the turns. Here, the burial and density regularizer terms favor the left-handed bundle, while the remainder of the terms reflecting secondary structure preferences, hydrogen bonding, and pair interactions favor the right-handed bundle. Again, the restraint energy difference is negligible. The combined total energy cannot differentiate between the two topological mirror images.

In some sense, the lack of discrimination power of the force field is similar to that experienced for other three- and four-helical bundles, such as protein A and 1hmd, which possess high internal symmetry.[2] In those cases and as here, it is not possible to confidently differentiate the topological mirror images, which are essentially isoenergetic within the resolution of the force field and protein representation.

In the context of simplified protein models, differences in the burial status of charged residues, particularly in turn regions, could result in a preference for one of two topological mirror images. The differences in surface coverage of the charged residues (Lys, Arg, Asp, Glu) between both images is plotted in Figure 5. Here, the surface coverage is calculated from an estimation of accessible surface area, which is based on the number of occupied lattice sites surrounding the residue of interest. This surface coverage has been calibrated to reproduce the average pairwise surface burial of residues, and is able to reproduce the continuous space result with an average error of 1.95% (unpublished results). Full surface coverage of

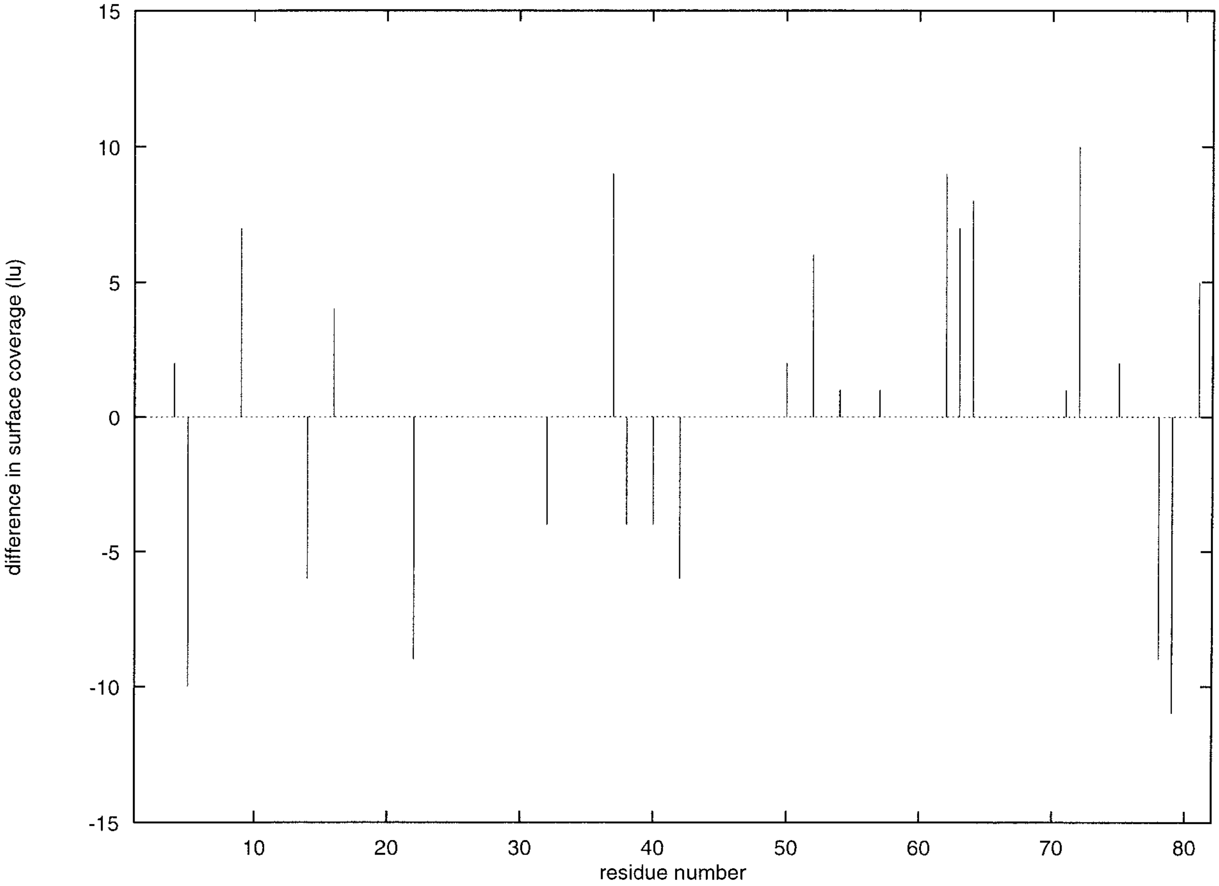

Fig. 5. Differences in the surface coverage of charged residues between the two topological mirror images shown in Figure 2. Full surface coverage of each residue amounts to 30 lattice points; thus, a difference of, for example, 7.5 lattice points corresponds to a difference in surface burial of 25%. Positive differences indicate that the corresponding residue in the right-handed topology is more buried than the corresponding residue in the left-handed topology.

each residue amounts to 30 lattice points; thus, differences in 7.5 lattice points correspond to differences in surface burial of 25%. Positive differences indicate that the residue in the right-handed topology is more buried than the corresponding residue in the left-handed topology. In general, the differences are small and are homogeneously distributed along the sequence. However, there is one continuous region of the sequence, between residues 50 and 75, and particularly between residues 50 and 63, in which the surface burial pattern strongly favors the left-handed topology over the right-handed one. It is of interest that the region between residues 50 and 63 corresponds to the turn between the long helices II and III. The results of this analysis are in agreement with an all-atom reconstruction of the lattice models (unpublished results), which confirms that residues 61 to 63 are buried in the all-atom model of the right-handed bundle. However, these differences are not captured by the statistical potential used in the simulations. Another interesting difference between the mirror images is related to their respective radii of gyration. The value in the left-handed topology is closer to the expected value than is the right-handed topology.

Expected Accuracy of the Prediction

In protein structure prediction, as important as the prediction itself is the expected accuracy that should be expected from the model, as well as the reliability of the prediction. Thus, it is of interest to give an estimate of the expected accuracy of the model, in terms of RMSD. KIX is an α-helical domain of 78 residues. It is possible to give an estimate of the accuracy by comparing the accuracy obtained previously by us with other helical proteins of similar size.[2] KIX seems to have the same number of residues and secondary structure content as the T0042 sequence of the recent CASP2 contest (for additional information about the CASP2 contest, see the web sites at URL http://iris4.carb.nist.gov/casp2/ and URL http://PredictionCenter.llnl.gov/). We blindly predicted the structure of T0042 as well. Comparison of the predicted model with the experimental structure indicated that the predicted model had a Cα RMSD of 5.6 with the experimental conformation. Furthermore, for other proteins, we have observed a significant correlation between the contact accuracy allowing a tolerance of ±2 residues ($\delta = 2$) and the RMSD of the predicted model with the experimental conformation. T0042 showed the expected behavior compared with the rest of the helical proteins. Three *seed* contacts were predicted for T0042, while four were used in the case of the KIX domain. As to the number of contacts, 24 contacts were used in the prediction of T0042 and 38 in the case of KIX. Since the numbers are comparable, a similar accuracy in contact prediction can be expected, the order of 58% at $\delta = 2$. On the other hand, the conformations explored during an isothermal calculation present a standard deviation of roughly 0.6 Å, which roughly matches the resolution of the lattice. Thus, the expected RMSD for the prediction of the KIX domain is between 5.0 and 6.2 Å. This is to be compared with the estimate of 12.9 Å that would be expected by a random prediction, using the model of Cohen and Sternberg.[23] As to the reliability of the prediction, only three-helix bundles were obtained from all assembly runs, but with about 50% population of each one of the two mirror images.

CONCLUSIONS

A blind prediction of the KIX domain of the protein CBP has been attempted using restraints derived from multiple sequence alignments coupled to Monte Carlo simulations. The predicted topology corresponds to a three-helix bundle. Because of possible errors in the tertiary restraints, the predicted structures may exhibit shifts in registration and distorted mutual orientations of pairs of secondary structural elements that, as a consequence, reduce the energy gap between the putative native conformation and alternative folds. Furthermore, we cannot determine which of the two topological mirror images is favored. This indicates that the interaction scheme needs improvement, and attempts to achieve this goal are now under way. For example, while the hydrophobic term can discriminate both topologies, the total force field energy cannot, suggesting that better balancing of terms is required. Some other parts of the procedure are also amenable to improvement as well, including refinements of the lattice representation itself. Contact map prediction is a key area, as the success of the prediction heavily depends on it. The combination of correlated mutations with threading appears to be able to select a coherent set of contacts adequate for forward folding, but improvements are certainly required to eliminate the chance of appearance of false positives, particularly for proteins larger than 100 amino acids.

ACKNOWLEDGMENTS

A.K. is an International Scholar of the Howard Hughes Medical Institute. We are grateful for access to the computational facilities of EMBL and the useful suggestions by the referees.

REFERENCES

1. Ortiz, A.R., Hu, W.-P., Kolinski, A., Skolnick, J. Method for low resolution prediction of small protein tertiary structure. In: “Pacific Symposium on Biocomputing ’97.” Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E. (eds.). Singapore: World Scientific, 1997:316–327.
2. Ortiz, A., Kolinski, A., Skolnick, J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J. Mol. Biol., in press, 1997.
3. Rost, B., Sander, C. Prediction of secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
4. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
5. Kolinski, A., Skolnick, J., Godzik, A., Hu, W.-P. A method for the prediction of surface “U”-turns and transglobular connections in small proteins. Proteins 27:290–308, 1997.
6. Godzik, A., Skolnick, J., Kolinski, A. A topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227: 227–238, 1992.
7. Hu, W.-P., Godzik, A., Skolnick, J. Sequence-structure specificity: How does an inverse folding approach work? Protein Eng. 10:317–331, 1997.
8. Skolnick, J., Kolinski, A., Ortiz, A.R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265:217–241, 1997.
9. Chrivia, J.C., Kwok, R.P., Lamb, N., Hagiwora, M., Goodman, R.M. Phosphorylated CREB binds specifically to the nuclear protein CBP. Nature 365:855–859, 1993.
10. Brindle, P., Linke, S., Montiminy, M. Protein kinase A dependent activator in transcription factor CREB reveals a new role for CREM repressors. Nature 364:821–824, 1993.
11. Goebel, U., Sander, C., Schneider, R., Valencia, A. Correlated mutations and residue contacts in proteins. Proteins 18:309–317, 1994.
12. Pearson, W.R., Lipman, D.J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85:2444–2448, 1988.
13. Sander, C., Schneider, R. Database of homology derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
14. McLachlan, A.D. Test for comparing related amino acid sequences. J. Mol. Biol. 61:409–424, 1971.
15. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
16. Kolinski, A., Skolnick, J. Determinants of secondary structure of polypeptide chains: Interplay between short rong and burial interactions. J. Chem. Phys. 107:953–964, 1997.
17. Skolnick, J., Jaroszewski, L., Kolinski, A., Godzik, A. Derivation and testing of pair potentials for protein folding: When is the quasichemical approximation correct? Protein Sci. 6:676–688, 1997.
18. Levitt, M., Greer, J. Automatic identification of secondary structure in globuar proteins. J. Mol. Biol. 114:181–293, 1977.
19. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 51:1087–1092, 1953.
20. Olmea, O., Valencia, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Folding & Design 2:S25–S32, 1997.
21. Pastore, A., Atkinson, R.A., Saudek, V., Williams, R.J.P. Topological mirror images in protein structure computation: An underestimated problem. Proteins 10:22–32, 1991.
22. Koradi, R., Billeter, M., Wuethrich, K. MOLMOL: A program for display and analysis of macromolecular structures. J. Mol. Graph. 14:51–55, 1996.
23. Cohen, F.E., Sternberg, M.J.E. On the prediction of protein structure: The significance of root-mean-square deviation. J. Mol. Biol. 138:321–333, 1980.