Gene 422 (2008) 47-51

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

A stochastic method for the reconstruction of protein structures from one-dimensional structural profiles

Katrin Wolff [a], Michele Vendruscolo [b], Markus Porto [a,*]

[a] Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstraße 6, 64289 Darmstadt, Germany
[b] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

ARTICLE INFO

Article history:
Available online 6 June 2008

Received by A. Bernardi

Keywords:
Protein folding
Structural profile
Tube model

ABSTRACT

We discuss a computational approach for reconstructing the native structures of proteins from the knowledge of a structural profile – the first eigenvector of the contact map of the native structure itself. The procedure consists in carrying out Monte Carlo simulations of a tube model of the protein structure with an energy bias towards the target structural profile. We present the reconstruction of two small proteins and address problems arising in the reconstruction of larger proteins. Our results indicate that an accurate physico-chemical energy function should be used in conjunction with the structural profile bias in order to achieve accurate reconstructions.

1. Introduction

The prediction of the structure of a protein from the knowledge of its amino acid sequence represents an ongoing challenge in biophysics and structural biology (see e.g. Moult et al., 2005). The most accurate results to date have been obtained by using approaches in which all-atom structures are constructed by optimising the assembly of fragments whose conformations are predicted at first (Simons et al. 1997). The prediction of the conformations of individual fragments is made by using bioinformatics tools for searching structural databases for similar amino acids sequences, and the subsequent assembly is carried out by using sophisticated all-atom force fields (see Bradley et al., 2005 and Schueler-Furman et al., 2005). Here, we take a different strategy by attempting the reconstruction of three-dimensional structure from a one-dimensional representation, or structural profile. This type of approach has been exploited in the alignment and comparison of protein structures as one-dimensional profiles can be handled more easily than three-dimensional structures (see e.g. Teichert et al., 2007). These results suggest that a one-dimensional profile can be used to help identify a three-dimensional protein structure and to reveal similarities between structures.

In the present work we consider a specific one-dimensional representation, the first eigenvector of the contact map of a protein structure (Porto et al., 2004). This profile is correlated to sequence hydrophobicity (Bastolla et al., 2005), and contact vector (Kabakcioglu et al., 2002) and can be predicted to good accuracy from the sequence (see e.g. Kinjo and Nishikawa, 2005). When using the contact map's principal eigenvector as structural profile it is in principle possible to reconstruct exactly the full structure from it, at least for single-domain proteins (Porto et al., 2004). Indeed, it has been shown by Porto et al. that the contact map can be reconstructed from its principal eigenvector by a deterministic algorithm. Moreover, it is also known that the contact map, even if incompletely known, is sufficient to determine the three-dimensional structure (Vendruscolo et al., 1997) thus resulting in the possibility of reconstructing three-dimensional structures from one-dimensional profiles. Chen et al. (2007) have shown that only about 70% of the information encoded in the contact map (equivalent to 1.5 constraints per residue) is necessary to reconstruct a protein structure to a C_α-RMSD of about 3 Å. Reconstruction from structural profiles corresponds to exactly one real-valued (and non-pairwise) constraint per residue. The first step of this prediction, i.e. the reconstruction of a contact map from its principal eigenvector, however, has been so far only possible for numerically exact profiles (Porto et al., 2004). The aim here is therefore to find a stochastic reconstruction scheme that is more robust to noise. At variance with the exact reconstruction procedure proposed by Porto et al., we attempt to reconstruct directly the three-dimensional structure from its one-dimensional representation, without going first through a contact map. In the general scheme that we envisage for predicting native structures of proteins from their amino acid sequences, the procedure that we describe in this work is eventually to be used for one-dimensional profiles that are predicted from amino acid sequences. Such profiles are inevitably affected by inaccuracies, or statistical noise, and are thus not amenable to the exact reconstruction procedure described by Porto et al. (2004). In the present work, however, we do not investigate directly the effect of the

Abbreviations: CASP, Critical assessment of structure prediction; PDB, Protein Databank; PE, Principal eigenvector; RMSD, Root mean square deviation.

* Corresponding author. TU Darmstadt, Institut für Festkörperphysik, Hochschulstraße 6, 64289 Darmstadt, Germany. Tel.: +49 6151 16 5395; fax: +49 6151 16 3681.
E-mail address: porto@fkp.tu-darmstadt.de (M. Porto).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.06.004

noise, but use the exact profiles obtained from actual protein structures to establish the general feasibility of a direct reconstruction.

We adopt here a computationally efficient structural model in which a protein conformation is represented as a tube of a shape that embeds the backbone of the protein and accounts in a coarse-grained manner for the excluded volume of the side chains (Hoang et al., 2004). The choice of the tube model not only enables to keep track of the amino acids' coordinates but also contains energy terms yielding realistic secondary structure motifs and protein-like behaviour.

2. Materials and methods

2.1. Tube model

Coarse-grained descriptions of protein structures allow the conformational space to be explored more efficiently than all-atom representations, and are therefore often very useful in computational approaches. In the most tractable models proteins are confined to a lattice (Sali et al., 1994). Valuable insight has been gained from this approach, but there are limits to how realistic it can be made. A promising new model has recently been proposed whose distinctive feature is that the protein backbone is assigned a finite thickness to account in an effective way for the volume occupied by the amino acid side chains (Hoang et al., 2004; Lezon et al., 2006; Banavar et al., 2006; Auer et al., 2007). The interactions considered include directional hydrogen bonding (with well depth e_{HB}), a local bending stiffness (defined by an energy penalty e_{S}), and pairwise attractive hydrophobic forces (with energy e_{W}). The protein is thus regarded as a uniform semi-flexible tube whose radial symmetry is broken by the restraints imposed by the hydrogen bonds. The excluded volume of the tube makes this model significantly different from other off-lattice coarse-grained models such as beads-on-strings, and also from Gō models because it includes no explicit energetic bias towards a predetermined structure. The energy of a protein in this model then is

$$E_{\mathrm{tube}} = \sum_i a(i{-}1,i,i+1)e_{\mathrm{S}} + \sum_{i<j} b(i,j)e_{\mathrm{W}} + \sum_{i<j} c(i{-}1,i,i+1,j{-}1,j,j+1)[e_{\mathrm{HB}}(i,j) + c(i,i+1,i+2,j,j+1,j+2)e_{\mathrm{coHB}}] \quad (1)$$

with the sums running over residues i. Functions a, b and c all are either 0 or 1 such that e_{S} becomes active if the angle at residue i is too tight, e_{W} if residues i and j are in contact, and e_{HB} if the conditions for hydrogen bonding are fulfilled. For $i>1$ and $j<N$ this means that the binormal vectors at i and j, as well as the vector connecting i and j, $\overrightarrow{r_{ij}}$, are all roughly parallel. For the first and last residue, where no binormal vectors can be defined, the definition of hydrogen bonds is altered to the constraint that $\overrightarrow{r_{ij}}$ make an angle between 70° and 110° with the extremal peptide links. The hydrogen bond energy $e_{\mathrm{HB}}(i,j)$ exists in two versions. For $j=i+3$ the hydrogen bond is considered local and $e_{\mathrm{HB}}(i,j)=-1$ (defining the energy scale for the model), for $j>i+3$ the bond is non-local and $e_{\mathrm{HB}}(i,j)=-0.7$. Local hydrogen bonds additionally require positive chirality $((\overrightarrow{r_{i,i+1}} \times \overrightarrow{r_{i+1,i+2}}) \cdot \overrightarrow{r_{i+2,i+3}} > 0)$. If residues i and j form a hydrogen bond and $i+1$ and $j+1$ do the same the structure gains energy for cooperative bonding, $e_{\mathrm{coHB}}=-0.3$. As there exist very good programs for side chain addition (see Canutescu et al., 2003 and references therein), successful backbone reconstruction is almost equivalent to complete reconstruction.

2.2. Contact map and principal eigenvector

A protein's contact map is a $N \times N$ symmetric matrix, where N is the number of amino acids, storing information about which amino acids are in contact (see Fig. 1). The matrix is binary with $C_{ij}=1$ if amino acids i and j are in contact and 0 otherwise. Two amino acids are defined as being in contact if the distance $x_{ij}=|\overrightarrow{r_{ij}}|$ between their C_α-atoms is less than a threshold value, for example $r_c=7.5$ Å,

$$C_{ij} = \begin{cases} 1 & x_{ij}<r_c \\ 0 & x_{ij}\geq r_c \end{cases}. \quad (2)$$

As a real symmetric matrix the contact map has N real eigenvalues. The structural profile then is the contact map's eigenvector to the largest eigenvalue (principal eigenvector or PE) and contains information about each amino acid's connectivity (see Fig. 2). Well connected amino acids that are in contact with many other residues have larger vector entries than those connected to fewer. In this context the correlation to hydrophobicity is also intuitively clear as residues with many contacts will be buried inside the protein fold as a high hydrophobicity requires (a discussion can be found in Bastolla et al., 2005). Note that the structural profile based on the PE, which is meaningful for single-domain folds, can be generalised to multi-domain folds (Teichert and Porto, 2006; Bastolla et al., 2006).

Alternative structural profiles include the principal eigenvector of a generalised contact map where contact is instead defined as

$$C_{ij} = \frac{1}{\exp\left(\frac{x_{ij}-x_0}{\Delta x}\right)+1} \quad (3)$$

and matrices are no longer binary. Profiles derived from this contact definition have the advantage of being a smooth function of the distance and thus exhibit smoother energy landscapes. For a plot of these contact functions see Fig. 3. Depending on the sharpness of the decline Δx, profiles of the second kind can be tuned to be more or less similar to the original definition and will be predictable from protein sequences with a similar accuracy.

2.3. Energy terms for Monte Carlo simulations

In order to carry out Monte Carlo simulations of the tube model we compare the profile of a candidate structure to the target profile obtained from the structure to be reconstructed, and define the energy as the sum of differences in the vector entries,

$$E_{\mathrm{PE}} = \sum_i \min(|v_i - t_i|, 0.25). \quad (4)$$

Here, a cutoff is introduced that limits the contribution of each vector entry to 0.25. For structures close to the target this cutoff makes

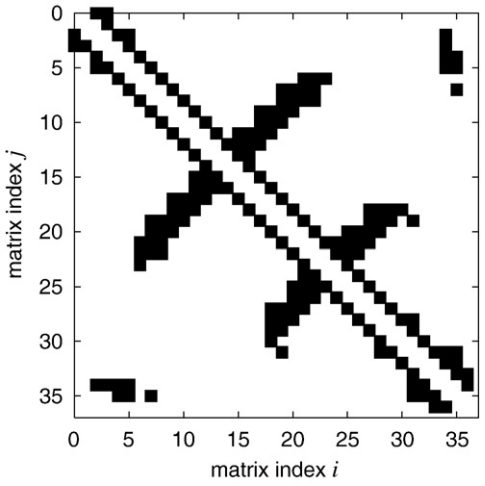

Fig. 1. Contact map of the native state of the FBP WW domain (PDB id. 1E0L), a small all-β protein ($N=37$). Black squares indicate contacts between amino acids. The native structure of this protein was completely recovered by the reconstruction procedure that we describe in this work.