Computational Biology and Chemistry 59 (2015) 1–7

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Machine Learnable Fold Space Representation based on Residue Cluster Classes

Ricardo Corral-Corral[a], Edgar Chavez[b], Gabriel Del Rio[a,*]

[a] Department of Biochemistry and Structural Biology, Instituto de Fisiologa Celular, Universidad Nacional Autónoma de México, México D. F., México
[b] Centro de Investigación Científica y de Educación Superior de Ensenada, México

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form 17 July 2015
Accepted 25 July 2015
Available online 30 July 2015

ABSTRACT

Motivation: Protein fold space is a conceptual framework where all possible protein folds exist and ideas about protein structure, function and evolution may be analyzed. Classification of protein folds in this space is commonly achieved by using similarity indexes and/or machine learning approaches, each with different limitations.
Results: We propose a method for constructing a compact vector space model of protein fold space by representing each protein structure by its residues local contacts. We developed an efficient method to statistically test for the separability of points in a space and showed that our protein fold space representation is learnable by any machine-learning algorithm.
Availability: An API is freely available at https://code.google.com/p/pyrcc/.

1. Introduction

All possible protein folds are assumed to occupy an abstract space referred to as fold space. This fold space has become a conceptual framework to unify ideas about protein structures with protein function and protein evolution (Cheng and Brooks, 2013). For instance, it is debated whether this space is discrete or continuous (Kolodny et al., 2006; Skolnick et al., 2009; Sadreyev et al., 2009). Relevant to our study is the common use of protein similarity measures (*e.g.*, root mean square deviation or RMSD) aimed to infer their proximities in this space (Minary and Levitt, 2008). In this case, the inference derived from such measurements assumes that the proximity of protein folds is the only relevant property to explain protein fold evolution and function.

Instead of focusing only on the proximity of protein folds, vector space models have been used to expand the protein fold space representation. In this space, each protein structure is represented in a fixed dimension space (*e.g.*, euclidean space) by a point (position vector); adjusting the positions of these vectors by approximating their relative distances to protein similarity measures may derive these position vectors. For example, using sequential structure alignment program scores between each pair of structures as a protein similarity measure (Orengo and Taylor, 1996), followed by multi dimensional scaling allowed the assignment of positions vectors representing each protein structure (Michie et al., 1996).

Using DALI as similarity measure, Holm (Holm and Sander, 1998) showed two-dimensional projections to explore protein neighbors in fold space. DALI has also been used as similarity measure to visualize class distribution and fold usages between two bacterial species (Hou et al., 2003) and to explore protein function assignment based on position on this fold space representation (Hou et al., 2005).

Fold space constructions based on protein structure similarities have two important limitations. First, the time needed to run a structural comparison for each pair of structures is restrictive: 25000 central processing unit hours were needed for calculating similarities between 1898 protein structures (Hou et al., 2005). Second, the position of each point depends heavily on the set of structures being analyzed, and in such case the inclusion of a single new structure can displace all previously assigned positions; thus, there are as many fold space representations as different protein structures data sets, even adopting a unique similarity score.

An alternative fold space representation may be built by assigning the position of the vector considering only the structural features of the protein it represents. In this way, a new structure can find its location in this fold space without altering the existing ones. One implementation of a vector space model is FragBag (Budowski-Tal et al., 2010), which represents protein structures in 400 dimensions; here, each component in the vector representing a protein fold corresponds to the number of occurrences of a

* Corresponding author. Tel.: +44 000 0000000; fax: +44 000 0000000
E-mail address: gdelrio@ifc.unam.mx (G. Del Rio).

http://dx.doi.org/10.1016/j.compbiolchem.2015.07.010
1476-9271/

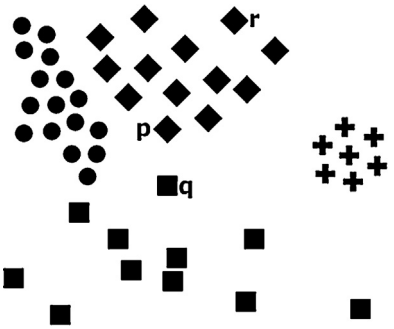

Fig. 1. Proximity in space does not always imply membership to a class. The figure illustrates cases where proximity in a space does not imply membership to a class. Cross points define a class where distances between every pair of its members should be less than the distance to any point belonging to other class, an assumption made when using similarity scores. An exception to this assumption is represented by the rhombus class members p and r, which distance is larger than for member class p to a square class member q. This is true for any distribution within the classes. To show this, circle class has a dense and regular distribution, rhombus class is regular and sparse, while square is very sparse and irregular. In any case, each class is clearly segregated. Please note that the different fold classes may be learned using machine-learning classifiers, but here we illustrate that using similarity measures as the only criterion to distinguish class membership may induce errors.

particular contiguous protein sequence fragment. Another approach uses knot invariants as values in each component for vector points in 30 dimensions (Rogen and Fain, 2003). In these cases, it is assumed that proximal protein structures should belong to the same structural class, assumption that is not necessarily correct as we will argue below.

Once a protein fold space construction is chosen, a metric distance induced by the space can be used as a measure of similarity and it is expected to be in agreement with direct structural measures (such as RMSD, GDT, TMscore, etc), but overcoming the problems noted above about these scores not being a metric (Sippl, 2008).

In such space, a given set of protein structures that are considered to have the same fold may be close in this space representation. Yet, it may occur that some proteins with different folds may be closer than proteins with the same fold (see Figure 1). Thus, confusion may be induced at distinguishing class membership in this fold space when only similarity measures are considered.

To address this problem, the boundaries between proteins with different folds may be obtained using empirical data and machine learning algorithms, which naturally segregate protein structures sharing common features. In these models it is also possible to evaluate the separability of this space independently of any machine-learning algorithm using a statistical test (Zighed et al., 2002). Therefore, it is possible to generate a protein fold space representation independent on the protein similarity measure used and to test for the separability of this space independently of any particular classification algorithm. This protein fold space may then be used to analyze protein structure-function relation and protein evolution without the limitations previously noted of current protein fold space construction approaches.

In this work, we propose a compact (low dimensional) fold space representation based on Residue Cluster Classes (RCCs), a Sperner family that includes all sets of residues in simultaneous contact. We also present an efficient computational method useful to test for the separability of this fold space representation. As a proof of principle, we analyzed the CATH classification and automatically detected conflicts in CATH. Furthermore, we show that our method improves state of the art protein structure neighbor retrieval methods. To facilitate the construction of protein folds represented by RCCs, we present an API available at https://code.google.com/p/pyrcc/.

2. Materials and Methods

2.1. Datasets

2.1.1. CATH datasets

CATHALL1 set includes all domains in CATH release v3.5 that were parsable with our API and consists of 168964 domains. CATHALL2 set includes all domains in CATH release v4.0 and contains 235858 domains. CATCHOP was obtained from a random sample of CATHALL1 considering only six domains per topology; topologies with less than six members were excluded rendering a total of 5220 domains (see supplemental Table S1 for a complete list).

2.1.2. SCOP datasets

The SCOP30 dataset was provided by the authors of ContactLib (Xuefeng Cui et al., 2014) and contains 3295 SCOP domains. SCOP30 contains 2639, 3232 and 3290 neighbours at SAS levels 20,35 and 50. Yet, only 2049, 2620 and 2722 domains have at least one neighbor in SAS 20, 35 and 50, respectively. The SCOPtrain1 is a random sample of 136300 SCOP 1.75B. SCOPtrain2 contains all 203025 domains from SCOP release v2.5. The SCOPtrainAUC includes 109310 SCOPtrain domains absent in SAS50 group, and only belonging to a class in SCOP30. If a class (at any level) contains more than 2000 domains, 2000 domains were chosen randomly to represent that class.

2.2. Construction of Residue Cluster Classes

2.2.1. Definitions

Residue Neighbourhood ($N_\epsilon(r)$). Let P be a protein with residues $R = r_1, r_2, \ldots, r_n$. The system τ_{prim} is defined as:

$$\tau_{prim} = \{\{r_i, r_j\} : \text{there exists a bond between } r_i \text{ and } r_j\}$$

Given a metric $d : R \times R \to [0, \infty)$ and a cut-off distance ϵ, the neighbourhood $N_\epsilon(r)$ of a residue r is given by:

$$N_\epsilon(r) = \{x \in R : \exists a \in A(x), b \in A(r); d(a, b) \leq \epsilon\}$$

Where $A(r)$ is the set of non-hydrogen atoms of residue r. Thus, $N_\epsilon(r)$ is the set of all residues near r, *i.e.*, they are at no more than a distance ϵ from r.

Residue Cluster (RC). A residue cluster on P is a subset $A \subseteq 2^R$ such that $A \subseteq N_\epsilon(a)$ for all $a \in A$. If $|A| = k$, then A is a k-Residue Cluster, ${}_kRC$

Residue Cluster Class (RCC). A class over a RC is defined by the primary structure on τ_{prim}. A pair of residues r_i, r_j are contiguous if $\{r_i, r_j\} \in \tau_{prim}$ and a set $\bar{s_L}$ of L residues form a segment in τ_{prim} if it contains $(L-1)$ contiguous residues. By convention, $\bar{s_1} = \{r\}$ for any $r \in R$.

Let be γ the family of all segments in τ_{prim}. C is a set cover of a k-Residue Cluster ${}_kRC$ if

$$C = \{\bar{s_{L_\alpha}} : \alpha \in A, \bar{s_{L_\alpha}} \in \gamma\}$$

such that

$${}_kRC = \bigcup\nolimits_{\alpha \in A} \bar{s_{L_\alpha}}$$

and

$$c_i \cap c_j = \emptyset, \forall c_i, c_j \in C$$

Therefore, $C = \bar{s_{L_1}}, \bar{s_{L_2}}, \ldots, \bar{s_{L_A}}$ is a covering if

$$\sum_{i=1}^{A} L_i = k$$