Applied Mathematics, 2014, 5, 2458-2463
Published Online September 2014 in SciRes. http://www.scirp.org/journal/am
http://dx.doi.org/10.4236/am.2014.516237

Discrete Differential Geometry of *n*-Simplices and Protein Structure Analysis

Naoto Morikawa

Genocript, Zama, Japan
Email: nmorika@genocript.com

Received 30 June 2014; revised 4 August 2014; accepted 17 August 2014

Abstract

This paper proposes a novel discrete differential geometry of *n*-simplices. It was originally developed for protein structure analysis. Unlike previous works, we consider connection between space-filling *n*-simplices. Using cones of an integer lattice, we introduce tangent bundle-like structure on a collection of *n*-simplices naturally. We have applied the mathematical framework to analysis of protein structures. In this paper, we propose a simple encoding method which translates the conformation of a protein backbone into a 16-valued sequence.

Keywords

Differential Geometry, *n*-Simplex, Discrete Mathematics, Protein Structure, Tetrahedron

1. Introduction

This paper proposes a novel discrete differential geometry of *n*-simplices, which is originally developed for protein structure analysis [1] [2]. Discrete differential geometry is the study of discrete equivalents of the geometric notions and methods of classical differential geometry [3] [4]. It mainly deals with polygonal curves and polyhedral surfaces whose properties are analogous to continuous counterparts, where the smooth theory is established as limit of the discrete theory.

On the other hand, we consider connection between space-filling *n*-simplices. We define gradient of *n*-simplices and obtain a flow of *n*-simplices by piling up *n*-cubes diagonally. Second derivative along a trajectory is given as a binary-valued sequence for any *n* (>1). As a result, we could encode the shape of *n*-dimensional objects if we approximate them by sweeping the occupied area with a trajectory of *n*-simplices.

Proteins are a sequence of amino acids linked by peptide bonds and fold into a unique three-dimensional structure in nature. Protein backbone structure is usually studied via manually-curated hierarchical classification [5] [6] but there also exist studies on differential geometric approach for protein structure analysis [7]-[11]. As

How to cite this paper: Morikawa, N. (2014) Discrete Differential Geometry of *n*-Simplices and Protein Structure Analysis. *Applied Mathematics*, **5**, 2458-2463. http://dx.doi.org/10.4236/am.2014.516237

for discrete differential geometry of protein backbones, proteins are usually represented as a polygonal chain, where curvature and torsion are defined at each vertex [7].

In our method, protein backbone structures are approximated by a trajectory of 3-simplices (tetrahedrons). Particularly we consider second derivative along a trajectory to encode local protein structures. Our method performs comparably with more sophisticated but more time-consuming methods which are specifically designed for protein structure analysis [12] [13]. In the following, we first describe the discrete differential geometry of *n*-simplices. Then, we apply the mathematical framework to analysis of protein structures and propose a simple encoding method which translates the conformation of a protein backbone into a 16-valued sequence.

2. Discrete Differential Geometry of *n*-Simplices

2.1. Basic Ideas

Recall that an *n*-simplex is an *n*-dimensional polytope which is the convex hull of its *n* + 1 vertices. As an introduction, we would consider the case of *n* = 2 before we give the definitions in the general case. In the case of *n* = 2, we obtain a flow of 2-simplices (triangles) by piling up unit cubes in the three-dimensional Euclidean space $\mathbb{R}^3$ as shown in **Figure 1(a)**.

First, cubes are pilled up in the direction of $(-1,-1,-1)$, where three upper faces of each unit cube are divided into two triangles by a diagonal line. Then, the diagonal lines on the faces of the cubes form a drawing on the surface of the "peaks and valleys" of cubes. By projecting the drawing onto a hyperplane that is perpendicular to $(1,1,1)$, a flow of triangles would be obtained. For example, the grey "slant" triangles on the surface specify the closed trajectory of the grey "flat" triangles on the hyperplane in **Figure 1(a)**.

2.2. Differential Structure

Because of convenience, we use monomials to represent coordinates of points. That is, point $(l_1, l_2, \cdots, l_n) \in \mathbb{R}^n$ is denoted by monomial $x_1^{l_1} x_2^{l_2} \cdots x_n^{l_n}$ of *n* indeterminates for integer *n* (*n* > 1).

First of all, we give the definition of "slant" and "flat" *n*-simplices. Let's consider *n*-cube in the *n*-dimensional Euclidean space $\mathbb{R}^n$. Note that the facets of *n*-cubes are $n-1$-dimensional unit cubes. To obtain "slant" *n*-simplices, we divide each of the *n* facets which contain origin $(0,0,\cdots,0)$ into $(n-1)(n-2)$ $n-1$-simplices along diagonal as follows.

Definition 1. For any integer *n* > 1, *n*-dimensional standard lattice L_n is the collection of all integer points of $\mathbb{R}^n$, *i.e.*,

$$L_n = \left\{ x_1^{l_1} x_2^{l_2} \cdots x_n^{l_n} \,\middle|\, l_i \in \mathbb{Z} \text{ for } 1 \le i \le n \right\}.$$

Definition 2. For any integer *n* > 1, the collection S_n of all slant *n*-simplices is defined by

$$S_n = \left\{ a\left[x_{\rho(1)} \cdots x_{\rho(n-1)} \right] \,\middle|\, a \in L_n,\ \rho \in Sym_n \right\},$$

where Sym_n is the *n*-th symmetric group and $a\left[x_{\rho(1)} \cdots x_{\rho(n-1)} \right]$ denotes the convex hull of *n* points $a_0 = a, a_1 = ax_{\rho(1)}, \cdots, a_{n-1} = ax_{\rho(1)} x_{\rho(2)} \cdots x_{\rho(n-1)} \in \mathbb{R}^n$ *i.e.*,

$$a\left[x_{\rho(1)} \cdots x_{\rho(n-1)} \right] = \left\{ \prod_{0 \le i < n} a_i^{\lambda_i} \,\middle|\, \lambda_i \in \mathbb{R}\ (0 \le i < n) \text{ s.t. } \lambda_i \ge 0 \text{ and } \sum_{0 \le i < n} \lambda_i = 1 \right\}.$$

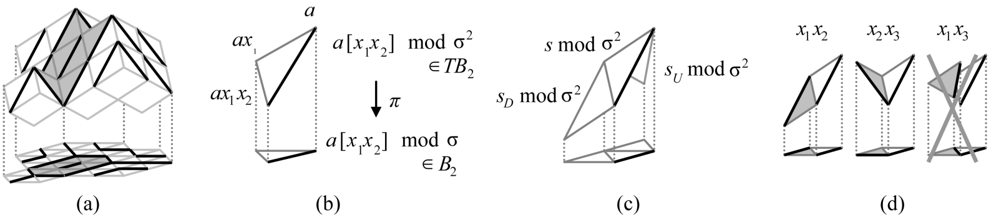


Figure 1. Discrete differential geometry of 2-simplices: (a) "Peaks and valleys" of cubes; (b) Tangent bundle-like structure; (c) Local trajectory; (d) Smoothness condition.

Definition 3. For any integer *n* > 1, the collection B_n of all flat *n*-simplices is defined as the quotient of S_n by "shift operator" σ on S_n, *i.e*,

$$B_n = S_n / \sigma ,$$

where $\sigma\left(a\left[x_{\rho(1)} \cdots x_{\rho(n-1)}\right]\right) = a x_{\rho(1)}\left[x_{\rho(2)} \cdots x_{\rho(n)}\right]$.

Definition 4. Tangent bundle-like structure TB_n is defined on B_n as the quotient of S_n by σ^n, *i.e.*,

$$\begin{cases} TB_n = S_n / \sigma^n , \\ \pi : TB_n \to B_n,\ \pi\left(s \bmod \sigma^n\right) = s \bmod \sigma. \end{cases}$$

For example, triangle $a\left[x_1 x_2\right]$ specifies element $a\left[x_1 x_2\right] \bmod \sigma^2$ of TB_n over element $a\left[x_1 x_2\right] \bmod \sigma$ of B_n (**Figure 1(b)**).

Definition 5. Gradient Ds of $s = a\left[x_{\rho(1)} \cdots x_{\rho(n-1)}\right] \in S_n$ is defined as a monomial of degree *n* − 1, *i.e.*,

$$Ds = x_{\rho(1)} x_{\rho(2)} \cdots x_{\rho(n-1)} .$$

For simplicity, we occasionally denote $x_{\rho(1)} \cdots x_{\rho(n-1)}$ by $e / x_{\rho(n)}$, where $e = x_1 x_2 \cdots x_n$. That is $Ds = e / x_{\rho(n)}$. Note that we could identify TB_n with $B_n \times \left\{e/x_1, e/x_2, \cdots, e/x_n\right\}$ by one-to-one correspondence

$$s \bmod \sigma^n \sim \left(s \bmod \sigma, Ds\right).$$

A gradient over a flat *n*-simplex specifies a local trajectory at the flat *n*-simplex as follows.

Definition 6. The local trajectory specified by $s \bmod \sigma^n \in TB_n$, where $s = a\left[x_{\rho(1)} \cdots x_{\rho(n-1)}\right] \in S_n$, is a collection of three adjacent flat *n*-simplices

$$\left\{s_D \bmod \sigma, s \bmod \sigma, s_U \bmod \sigma\right\},$$

where $s_D = a x_{\rho(1)}\left[x_{\rho(2)} \cdots x_{\rho(n-1)} x_{\rho(1)}\right]$ and $s_U = a\left[x_{\rho(1)} \cdots x_{\rho(n-2)} x_{\rho(n)}\right]$.

For example, $\left[x_1 x_2\right] \bmod \sigma^2 \in TB_2$ specifies local trajectory $\left\{x_1\left[x_2 x_1\right] \bmod \sigma, \left[x_1 x_2\right] \bmod \sigma, \left[x_1 x_3\right] \bmod \sigma\right\}$ at $\left[x_1 x_2\right] \bmod \sigma \in B_2$ (**Figure 1(c)**). We would obtain a flow on B_n by patching these local trajectories together.

To define the "second derivative" along a trajectory, we would impose a kind of "smoothness condition" on local trajectories.

Definition 7. (Smoothness condition). Let Γ be a section of TB_n on $\left\{s_D \bmod \sigma, s \bmod \sigma, s_U \bmod \sigma\right\} \subset B_n$ where $s = a\left[x_{\rho(1)} \cdots x_{\rho(n-1)}\right] \in S_n$. Suppose that $D\left[\Gamma\left(s \bmod \sigma\right)\right] = e / x_{\rho(n)}$. Then, we impose the following conditions on the local trajectory:

$$\begin{cases} D\left[\Gamma\left(s_D \bmod \sigma\right)\right] = e / x_{\rho(n)} \text{ or } e / x_{\rho(1)} , \\ D\left[\Gamma\left(s_U \bmod \sigma\right)\right] = e / x_{\rho(n)} \text{ or } e / x_{\rho(n-1)} . \end{cases}$$

Remark 8. For any two consecutive *n*-simplices $\left\{t_1, t_2\right\}$ on a trajectory, there exist $a \in L_n$ and $\rho \in Sym_n$ s.t. $t_1 = a\left[x_{\rho(1)} \cdots x_{\rho(n-2)} x_{\rho(n-1)}\right] \bmod \sigma$ and $t_2 = a\left[x_{\rho(1)} \cdots x_{\rho(n-2)} x_{\rho(n)}\right] \bmod \sigma$. Monomial $x_{\rho(1)} \cdots x_{\rho(n-2)}$ is uniquely determined by $\left\{t_1, t_2\right\}$ and is included in both $D\left[\Gamma\left(t_1\right)\right]$ and $D\left[\Gamma\left(t_2\right)\right]$ for any section Γ of TB_n on $\left\{t_1, t_2\right\}$. That is, $x_{\rho(1)} \cdots x_{\rho(n-2)}$ corresponds to the contact surface between two consecutive slant *n*-simplices.

As an example, let's consider the case of *n* = 2 shown in **Figure 1(d)**, where the gradient at current triangle $a\left[x_1 x_2\right] \bmod \sigma$ is $x_1 x_2$. Then, the gradient at next triangle $a x_1\left[x_2 x_1\right] \bmod \sigma$ could assume either $x_1 x_2$ or $x_2 x_3$. Otherwise, we couldn't connect the two consecutive slant triangles over the trajectory "smoothly" as shown in the figure.

2.3. Tangent Cone and Section of *TB_n*

Now we give the definition of the "peaks and valleys" of *n*-simplices (**Figure 1(a)**).

Definition 9. For $A \subset L_n$, tangent cone $\text{Cone}\, A$ of L_n is defined as follows:

$$\text{Cone}\, A = \left\{ p x_1^{l_1} x_2^{l_2} \cdots x_n^{l_n} \,\middle|\, p \in A \text{ and } l_i \geq 0 \ (1 \leq i \leq n) \right\}.$$

Definition 10. For tangent cone $w = \text{Cone}\, A \ (A \subset L_n)$, boundary surfaces $d_S w$ is defined by

$$d_S w = \left\{ a\left[x_{\rho(1)} \cdots x_{\rho(n-1)} \right] \in S_n \,\middle|\, l_w(a_i) = 0 \ (0 \leq i < n) \right\},$$

where $a_0 = a, a_1 = a x_{\rho(1)}, \cdots, a_{n-1} = a x_{\rho(1)} x_{\rho(2)} \cdots x_{\rho(n-1)} \in \mathbb{R}^n$ and, for $z \in L_n$,

$$l_w(z) = \max_{p \in w} \left\{ \min\left\{ l_1, l_2, \cdots, l_n \,\middle|\, l_i \in \mathbb{Z} \ (1 \leq i \leq n) \text{ s.t. } x_1^{l_1} x_2^{l_2} \cdots x_n^{l_n} = z/p \right\} \right\}.$$

Then, $d_S w$ specifies a unique slant *n*-simplex over each $t \in B_n$ and we obtain a section of S_n on B_n.

Definition 11. Γ_w is the section of S_n on B_n induced by tangent cone w, *i.e.*, for $t \in B_n$,

$$\Gamma_w(t) = s \in d_S w \text{ s.t. } t = s \bmod \sigma.$$

Note that tangent cone w induces a section of TB_n on B_n by $D\Gamma_w : B_n \to \{e/x_1, e/x_2, \cdots, e/x_n\}$. Patching the local trajectories specified by $D\Gamma_w$ together, we would obtain a flow on B_n. As an example, let's consider the "peaks and valleys" shown in **Figure 1(a)**, which is induced by $w = \text{Cone}\{1, x_1 x_2^{-1}, x_1^2 x_2 x_3^{-1}\}$.

Let's start from triangle $[x_1 x_2] \bmod \sigma$ (grey) and move downward (**Figure 2**): $t[0] = [x_1 x_2] \bmod \sigma$ and $D\Gamma_w(t[0]) = x_1 x_2$. $D\Gamma_w(t[0])$ specifies local trajectory $\{x_1[x_2 x_1] \bmod \sigma, [x_1 x_2] \bmod \sigma, [x_1 x_3] \bmod \sigma\}$ at $t[0]$. Since we move downward, next triangle $t[1]$ is $x_1[x_2 x_1] \bmod \sigma$ and we obtain $D\Gamma_w(t[1]) = x_1 x_2$. Then, $D\Gamma_w(t[1])$ specifies local trajectory $\{x_1 x_2 [x_1 x_2] \bmod \sigma, x_1[x_2 x_1] \bmod \sigma, [x_1 x_2] \bmod \sigma\}$ at $t[1]$ and next traingle $t[2]$ is $x_1 x_2 [x_1 x_2] \bmod \sigma$. Continuing the process, we obtain a closed trajectory of length 10.

Finally, we consider variation of gradient, *i.e.*, "second derivative", along a trajectory. Thanks for the smoothness condition, variation of gradient along a trajectory could be specified as a binary valued sequence.

Definition 12. Let $\{t[i]\} \subset B_n$ be a trajectory induced by $D\Gamma_w$ for tangent cone w. Then, "second derivative" $D^2\Gamma_w$ of Γ_w along $\{t[i]\}$ is defined as a $\{U, D\}$-valued function:

$$D^2\Gamma_w(t[i+1]) = \begin{cases} D^2\Gamma_w(t[i]) & \text{if } D\Gamma_w(t[i+1]) = D\Gamma_w(t[i]), \\ -D^2\Gamma_w(t[i]) & \text{else}, \end{cases}$$

where $-D = U$ and $-U = D$.

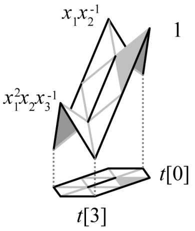
Then, we could encode the conformation of a trajectory by the second derivative along the trajectory. As an example, let's consider the trajectory of **Figure 2** again. First, set any initial value: $D^2\Gamma_w(t[0]) = D$. Then, since the first two triangles $t[0]$ and $t[1]$ have the same gradient, $D^2\Gamma_w(t[1]) = D$. The value of the second derivative is *D* until $t[3]$, where it is changed to *U* because the gradient of $t[2]$ is different from that of $t[3]$. Continuing the process, we obtain a binary sequence of length 10, *DDDUDUUUDU*, which describes the shape swept by the trajectory of triangles.

3. Encoding of Protein Backbone Structure

In the case of *n* = 3, we obtain a flow of 3-simplices (tetrahedrons), which is used for protein structure analysis. In this section we propose a simple encoding method which translates the conformation of a protein backbone into a sequence of letters from a 16-letter alphabet (called D^2 codes), using the second derivative along trajectories of tetrahedrons.

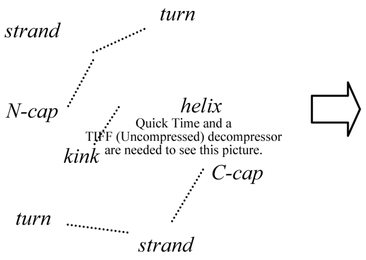
First, we consider all the fragments of five amino-acids occurred in a protein. Each fragment is approximated by a tetrahedron sequence of length five, where we permit translation and rotation during the process to absorb irregularity inherent in actual protein structures.

Next, we compute the second derivative along the tetrahedron sequences to obtain binary-valued sequences of length five. We assign the binary-valued sequences, which are denoted as a base-32 number, to the center amino-acid of the corresponding fragment. For example, *DDDUD* is denoted by "2", *DUDDU* is denoted by "9", *DUDUD* is denoted by "*A*", and so on.



Trajectory	***Gradient***	***2nd Derivative***
$t[0]=[x_1x_2] \bmod \sigma$	$D\Gamma_w(t[0])=x_1x_2$	$D^2\Gamma_w(t[0])=D$
$t[1]=x_1[x_2x_1] \bmod \sigma$	$D\Gamma_w(t[1])=x_1x_2$	$D^2\Gamma_w(t[1])=D$
$t[2]=x_1x_2[x_1x_2] \bmod \sigma$	$D\Gamma_w(t[2])=x_1x_2$	$D^2\Gamma_w(t[2])=D$
$t[3]=x_1^2x_2x_3^{-1}[x_3x_2] \bmod \sigma$	$D\Gamma_w(t[3])=x_2x_3$	$D^2\Gamma_w(t[3])=U$
…	…	…

Figure 2. Closed trajectory of 2-simplices induced by $\mathrm{Cone}\left\{1, x_1x_2^{-1}, x_1^2x_2x_3^{-1}\right\}$.


strand
turn
N-cap
helix
Quick Time and a
TIFF (Uncompressed) decompressor
are needed to see this picture.
kink
C-cap
turn
strand

000RQAAAAAAAHAAAAAAAAAAAB0R00000

strand	⇔	*0* (=*DDDDD*)
turn	⇔	*R* (=*UUDUU*)
N-cap	⇔	*Q* (=*UUDUD*)
helix	⇔	*A* (=*DUDUD*)
C-cap	⇔	*B* (=*DUDUU*)
kink	⇔	*H* (=*UDDDU*)

Figure 3. D^2-encoding of a protein (transferase 1RKL).

Then, we obtain a one-dimensional representation of protein backbone structure by arranging the base-32 numbers in the order the corresponding amino-acids appear in the protein. See [1] for detailed description of the algorithm.

Figure 3 shows an example of D^2-encoding of a protein. As you see, our method captures successfully not only recurring structural features of the protein (strand, turn, caps, helix), but also distortions (such as kink) as well.

4. Discussion

In this paper, we first describe the discrete differential geometry of *n*-simplices. Then, we apply the mathematical framework to analysis of protein structures and propose a simple encoding method which translates the conformation of a protein backbone into a 16-valued sequence.

Unlike previous works, our version of discrete differential geometry studies connection between space-filling *n*-simplices. Considering cones of an integer lattice, we have introduced tangent bundle-like structure on *n*-simplices naturally. On notable consequence is the smoothness condition, *i.e.*, restriction on variation of gradient along a trajectory. In particular, we could encode the shape of *n*-dimensional objects if we approximate them by sweeping the occupied area with a trajectory of *n*-simplices.

As for protein structure analysis, since we do not use clustering analysis to encode local structures, our approach not only provides a intuitively understandable description of protein structures, but also covers wide varieties of distortions. Our method performs comparably with more sophisticated but more time-consuming methods which are specifically designed for protein structure analysis. In SHREC'10 Protein Model Classification we achieved results comparable to more sophisticated methods, using the length of the longest common subsequence as the measure of structural similarity [12]. At homology level of CATH95 data set, our method performs best among all the individual classifiers considered in [13].

References

[1] Morikawa, N. (2007) Discrete Differential Geometry of Tetrahedrons and Encoding of Local Protein Structure. arXiv: math.CO/0710.4596.

[2] Morikawa, N. (2011) A Novel Method for Identification of Local Conformational Changes in Proteins. arXiv: q-bio. BM/1110.6250.

[3] Bobenko, A.I. and Suris, Yu.B. (2008) Discrete Differential Geometry. Integrable Structure. *Graduate Studies in Mathematics*, **98**, 404 p. arXiv:math/0504358.

[4] Meyer, M., Desbrun, M., Schroder, P. and Barr, A.H. (2003) Discrete Differential-Geometry Operators for Triangulated 2-Manifolds. In: Hege, H.-C. and Polthier, K., Eds., *Visualization and Mathematics III*, Springer-Verlag, Berlin, 35-58. http://dx.doi.org/10.1007/978-3-662-05105-4_2

[5] Sillitoe, I., *et al*. (2013) New Functional Families (FunFams) in CATH to Improve the Mapping of Conserved Functional Sites to 3D Structures. *Nucleic Acids Research*, **41**, D490-D498. http://dx.doi.org/10.1093/nar/gks1211

[6] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology*, **247**, 536-540. http://dx.doi.org/10.1016/S0022-2836(05)80134-2

[7] Rackovsky, S. and Scheraga, H.A. (1978) Differential Geometry and Polymer Conformation. 1. Comparison of Protein Conformations. *Macromolecules*, **11**, 1168-1174. http://dx.doi.org/10.1021/ma60066a020

[8] Louie, A.H. and Somorjai, R.L. (1982) Differential Geometry of Proteins: A Structural and Dynamical Representation of Patterns. *Journal of Theoretical Biology*, **98**, 189-209. http://dx.doi.org/10.1016/0022-5193(82)90258-2

[9] Montalvao, R.W., Smith, R.E., Lovell, S.C. and Blundell, T.L. (2005) CHORAL: A Differential Geometry Approach to the Prediction of the Cores of Protein Structures. *Bioinformatics*, **21**, 3719-3725. http://dx.doi.org/10.1093/bioinformatics/bti595

[10] Gorielyn, A., Hausrath, A. and Neukirch, S. (2008) The Differential Geometry of Proteins and Its Applications to Structure Determination. *Biophysical Reviews and Letters*, **3**, 77-101. http://dx.doi.org/10.1142/S1793048008000629

[11] Hu, S., Lundgren, M. and Niemi, A.J. (2011) The Discrete Frenet Frame, Inflection Point Solitons and Curve Visualization with Applications to Folded Proteins. arXiv:q-bio.BM/1102.5658.

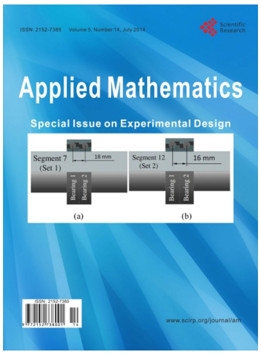
[12] Mavridis, L., *et al*. (2010) SHREC'10 Track: Protein Model Classification. *Proceedings of Eurographics Workshop on* 3*D Object Retrieval*, 117-124.

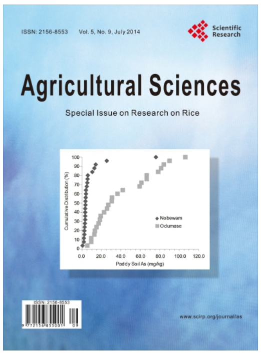
[13] Boujenfa, K. and Limam, M. (2012) Consensus Decision for Protein Structure Classification. *Journal of Intelligent Learning Systems and Applications*, **4**, 216-222. http://dx.doi.org/10.4236/jilsa.2012.43022

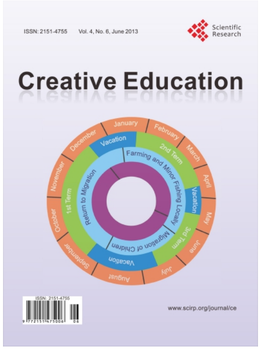
Advances in Bioscience and Biotechnology

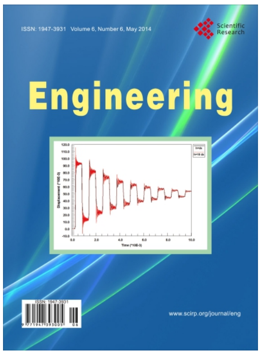
American Journal of Analytical Chemistry

American Journal of Plant Sciences

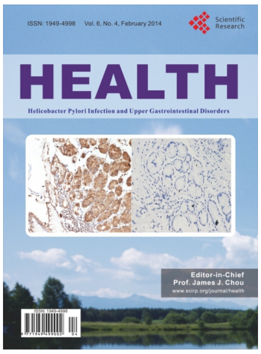

Applied Mathematics


Agricultural Sciences


Creative Education

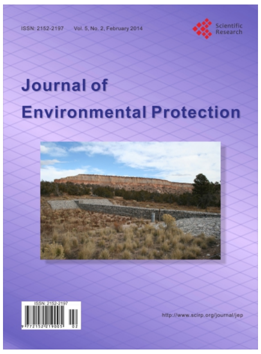

Engineering

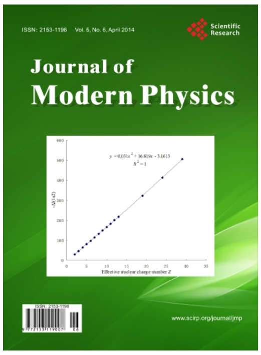
Food and Nutrition Sciences


HEALTH

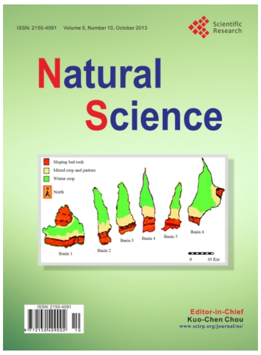
Journal of Computer and Communications

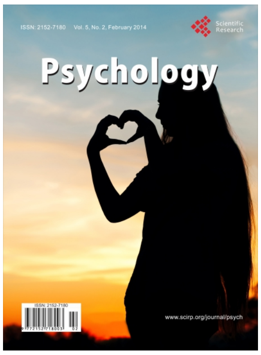
Journal of Cancer Therapy


Journal of Environmental Protection


Journal of Modern Physics

Modern Economy


Natural Science


Psychology