Submitted to *Symmetry*. Pages 1 - 7.

*Article*

# Symmetries Explain Why We Observe Alpha-Helices, Beta-Sheets, and Beta-Barrels in Protein Structure: Towards Further Development of Gromov's Ideas

**Jaime Nava and Vladik Kreinovich**$^{\star}$

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso TX 79968, USA

$^{\star}$ Author to whom correspondence should be addressed; vladik@utep.edu, Tel. +1-915-747-6951, Fax +1-915-747-5030.

*Version July 17, 2011 submitted to* Symmetry. *Typeset by LaTeX using class file mdpi.cls*

**Abstract:** Protein structure is invariably connected to protein function. There are two important secondary structure elements: alpha helices and beta-sheets – which also come in a shape of beta-barrels. The actual shapes of these structures can be complicated, but in the first approximation, they are usually approximated by spirals, planes, and cylinders. In this paper, following Misha Gromov's ideas, we use natural symmetries to show that, under reasonable assumptions, these sets are indeed the best approximating families for secondary structures.

**Keywords:** symmetries; secondary protein structures; alpha-helices; beta-sheets; beta-barrels

## 1. Introduction

Proteins are biological polymers that perform most of the life's function. A single chain polymer (protein) is folded in such a way that forms local substructures called secondary structure elements. In order to study the structure and function of proteins it is extremely important to have a good geometrical description of the proteins structure. There are two important secondary structure elements: alpha helices and beta-sheets. A part of the protein structure where different fragments of the polypeptide align next to each other in extended conformation forming a surface-like feature defines a secondary structure called a *beta pleated sheet*, or, for short, a *beta-sheet*; see, e.g., [1,6].

Beta-sheets are coming in many forms and shapes. In some cases, we have a cylinder-like structure called a *beta-barrel* that is "closed" in one dimension and "open" in the other, but in most cases, we have a surface that is open in both directions.

The actual shapes of the alpha-helices, beta-sheets, and beta-barrels can be complicated. In the first approximation, they are usually approximated by cylindrical spirals, planes, and cylinders. In this paper, following Misha Gromov's ides [5], we use symmetries to show that under reasonable assumptions, these empirically observed shapes are indeed the best families of simple approximating sets.

Thus, symmetries indeed explain why the secondary protein structures consists of alpha-helices, beta-sheets, and beta-barrels.

## 2. Analysis of the Problem

Of course, the more parameters we allow, the better the approximation. So, the question of selecting the best approximating family of sets can be reformulated as follows: for a given number of parameters (i.e., for a given dimension of approximating family of sets), which is the best family?

When we say "the best", we mean that on the set of all appropriate families, there is a relation $\succeq$ describing which family is better or equal in quality. This relation must be transitive (if $A$ is better than $B$, and $B$ is better than $C$, then $A$ is better than $C$). This relation is not necessarily asymmetric, because we can have two approximating families of the same quality. However, we would like to require that this relation be *final* in the sense that it should define a unique *best* family $A_{\text{opt}}$ (i.e., the unique family for which $\forall B\,(A_{\text{opt}} \succeq B)$. Indeed:

- If none of the families is the best, then this criterion is of no use, so there should be *at least one* optimal family.
- If *several* different families are equally best, then we can use this ambiguity to optimize something else: e.g., if we have two families with the same approximating quality, then we choose the one which is easier to compute. As a result, the original criterion was not final: we get a new criterion ($A \succeq_{\text{new}} B$ if either $A$ gives a better approximation, or if $A \sim_{\text{old}} B$ and $A$ is easier to compute), for which the class of optimal families is narrower. We can repeat this procedure until we get a final criterion for which there is only one optimal family.

It is reasonable to require that the relation $A \succeq B$ should be invariance relative to natural geometric symmetries, i.e., shift- and rotation-invariant.

These requirements sounds reasonable but weak. We will show, however, that they are sufficient to find the optimal families.

*Comment.* Our explanation is similar to the symmetry-based explanation of the shapes of celestial bodies presented in [2–4,7].

## 3. Definitions and the Main Mathematical Result

Our goal is to choose the best finite-parametric family of sets. To formulate this problem precisely, we must formalize what a finite-parametric family is and what it means for a family to be optimal. In accordance with the above analysis of the problem, in both formalizations will use natural symmetries.

So, we will first formulate how symmetries can be defined for families of sets, then what it means for a family of sets to be finite-dimensional, and finally, how to describe an optimality criterion.

**Definition 1.** *Let $g : M \to M$ be a 1-1-transformation of a set $M$, and let $A$ be a family of subsets of $M$. For each set $X \in A$, we define the result $g(X)$ of applying this transformation $g$ to the set $X$ as $\{g(x) \,|\, x \in X\}$, and we define the result $g(A)$ of applying the transformation $g$ to the family $A$ as the family $\{g(X) \,|\, X \in A\}$.*

**Definition 2.** *Let $M$ be a smooth manifold. A group $G$ of transformations $M \to M$ is called a* Lie transformation group*, if $G$ is endowed with a structure of a smooth manifold for which the mapping $g, a \to g(a)$ from $G \times M$ to $M$ is smooth.*

We want to define $r$-parametric families sets in such a way that symmetries from $G$ would be computable based on parameters. Formally:

**Definition 3.** *Let $M$ and $N$ be smooth manifolds.*

- *By a* multi-valued function *$F : M \to N$ we mean a function that maps each $m \in M$ into a discrete set $F(m) \subseteq N$.*
- *We say that a multi-valued function is* smooth *if for every point $m_0 \in M$ and for every value $f_0 \in F(m)$, there exists an open neighborhood $U$ of $m_0$ and a smooth function $f : U \to N$ for which $f(m_0) = f_0$ and for every $m \in U$, $f(m) \subseteq F(m)$.*

**Definition 4.** *Let $G$ be a Lie transformation group on a smooth manifold $M$.*

- *We say that a class $A$ of closed subsets of $M$ is $G$-*invariant *if for every set $X \in A$, and for every transformation $g \in G$, the set $g(X)$ also belongs to the class.*
- *If $A$ is a $G$-invariant class, then we say that $A$ is a* finitely parametric family of sets *if there exist:*
    - *a (finite-dimensional) smooth manifold $V$;*
    - *a mapping $s$ that maps each element $v \in V$ into a set $s(v) \subseteq M$; and*
    - *a smooth multi-valued function $\Pi : G \times V \to V$*

  *such that:*

    - *the class of all sets $s(v)$ that corresponds to different $v \in V$ coincides with $A$, and*
    - *for every $v \in V$, for every transformation $g \in G$, and for every $\pi \in \Pi(g, v)$, the set $s(\pi)$ (that corresponds to $\pi$) is equal to the result $g(s(v))$ of applying the transformation $g$ to the set $s(v)$ (that corresponds to $v$).*
- *Let $r > 0$ be an integer. We say that a class of sets $B$ is a $r$-*parametric class *of sets if there exists a finite-dimensional family of sets $A$ defined by a triple $(V, s, \Pi)$ for which $B$ consists of all the sets $s(v)$ with $v$ from some $r$-dimensional sub-manifold $W \subseteq V$.*

**Definition 5.** *Let $\mathcal{A}$ be a set, and let $G$ be a group of transformations defined on $\mathcal{A}$.*

- *By an* optimality criterion*, we mean a* pre-ordering *(i.e., a transitive reflexive relation)* $\preceq$ *on the set* $\mathcal{A}$*.*

- *An optimality criterion is called* $G$-invariant *if for all* $g \in G$*, and for all* $A, B \in \mathcal{A}$*,* $A \preceq B$ *implies* $g(A) \preceq g(B)$*.*

- *An optimality criterion is called* final *if there exists one and only one element* $A \in \mathcal{A}$ *that is preferable to all the others, i.e., for which* $B \preceq A$ *for all* $B \neq A$*.*

**Proposition.** *Let* $M$ *be a manifold, let* $G$ *be a* $d$*-dimensional Lie transformation group on* $M$*, and let* $\preceq$ *be a* $G$*-invariant and final optimality criterion on the class* $\mathcal{A}$ *of all* $r$*-parametric families of sets from* $M$*,* $r < d$*. Then:*

- *the optimal family* $A_{\text{opt}}$ *is* $G$*-invariant; and*

- *each set* $X$ *from the optimal family is a union of orbits of* $\geq (d - r)$*-dimensional subgroups of the group* $G$*.*

*Comment.* For readers' convenience, the proof of the Proposition is placed in the special (last) section.

## 4. Resulting Geometric Shapes

In our case, the natural group of symmetries $G$ is generated by shifts and rotations. So, to apply the above Proposition to the geometry of protein structures, we must describe all orbits of subgroups of this groups $G$.

In the applications to the geometry of a molecule, we only considered connected *continuous* subgroups $G_0 \subseteq G$: since connected continuous subgroups explain connected shapes.

Let us start with 1-D orbits. A 1-D orbit is an orbit of a 1-D subgroup. This subgroup is uniquely determined by its "infinitesimal" element, i.e., by the corresponding element of the Lie algebra of the group $G$. This Lie algebra if easy to describe. For each of its elements, the corresponding differential equation (that describes the orbit) is reasonably easy to solve.

2-D forms are orbits of $\geq$ 2-D subgroups, so, they can be enumerated by combining two 1-D subgroups.

*Comment.* An alternative (slightly more geometric) way of describing 1-D orbits is to take into consideration that an orbit, just like any other curve in a 3-D space, is uniquely determined by its curvature $\kappa_1(s)$ and torsion $\kappa_2(s)$, where $s$ is the arc length measured from some fixed point. The fact that this curve is an orbit of a 1-D group means that for every two points $x$ and $x'$ on this curve, there exists a transformation $g \in G$ that maps $x$ into $x'$. Shifts and rotations do not change $\kappa_i$, they may only shift $s$ (to $s + s_0$). This means that the values of $\kappa_i$ are constant. Taking constant $\kappa_i$, we get differential equations, whose solution leads to the desired 1-D orbits.

The resulting description of 0-, 1-, and 2-dimensional orbits of connected subgroups $G_a$ of the group $G$ is as follows:

0: The only 0-dimensional orbit is a *point*.

1: A generic 1-dimensional orbit is a *cylindrical spiral*, which is described (in appropriate coordinates) by the equations $z = k \cdot \phi$, $\rho = R_0$. Its limit cases are:

- a *circle* ($z = 0$, $\rho = R_0$);
- a *semi-line* (*ray*);
- a *straight line*.

2: Possible 2-D orbits include:

- a *plane*;
- a *semi-plane*;
- a *sphere*; and
- a *circular cylinder*.

Bounded shapes like a point, a circle, or a sphere do occur in chemistry, but, due to their boundedness, they usually (approximately) describe the shapes of relatively small molecules like benzenes, fullerenes, etc. We are interested in relatively large molecules like proteins, so it is reasonable to only consider unbounded shapes. With this restriction, we end up with the following shapes:

- a cylindrical spiral (with a straight line as its limit case);
- a plane (or a part of the plane), and
- a cylinder.

These shapes correspond exactly to alpha-helices, beta-sheets, and beta-barrels that we observe in proteins. Thus, the symmetries indeed explain the observed protein shapes.

*Comment.* As we have mentioned earlier, spirals, planes, and cylinders are only the first approximation to the actual shape of protein structures. For example, it has been empirically found that for beta-sheets and beta-barrels, general hyperbolic (quadratic) surfaces provide a good second approximation; see, e.g., [8]. It is worth mentioning that the empirical fact that quadratic models provide the best second approximation can also be theoretical explained by using symmetries [9].

## 5. Possible Physical Meaning

We have provided a somewhat mathematical explanation for the shapes, but this explanation can be also reformulated in more physical terms. In the beginning, protein generation starts with a uniform medium, in which the distribution is homogeneous and isotropic. In mathematical terms, the initial distribution of matter is invariant w.r.t. arbitrary shifts and rotations.

The equations that describe the physical forces that are behind the corresponding chemical reactions are invariant w.r.t. arbitrary shifts and rotations. In other words, these interactions are *invariant* w.r.t. our group $G$. The *initial distribution* was *invariant* w.r.t. $G$; the *evolution equations are* also *invariant*; hence, at first glance, we should get a $G$-invariant distribution of for all moments of time.

In reality, we do not see such a homogeneous distribution – because this highly symmetric distribution is known to be *unstable*. As a result, an arbitrarily small perturbations cause drastic changes in the matter distribution: matter concentrates in some areas, and shapes are formed. In physics, such symmetry violation is called *spontaneous*.

In principle, it is possible to have a perturbation that changes the initial highly symmetric state into a state with no symmetries at all, but statistical physics teaches us that it is much more probable to have a gradual symmetry violation: first, some of the symmetries are violated, while some still remain; then, some other symmetries are violated, etc.

Similarly, a (highly organized) solid body normally goes through a (somewhat organized) liquid phase before it reaches a (completely disorganized) gas phase.

If a certain perturbation concentrates matter, among other points, at some point $a$, then, due to invariance, for every transformation $g \in G'$, we will observe a similar concentration at the point $g(a)$. Therefore, the shape of the resulting concentration contains, with every point $a$, the entire *orbit* $G'(a) = \{g(a) \,|\, g \in G'\}$ of the group $G'$. Hence, the resulting *shape consists of* one or several *orbits* of a group $G'$. This is exactly the conclusion we came up with before, but now we have a physical explanation for it.

## 6. Proof of Proposition

Since the criterion $\preceq$ is final, there exists one and only one optimal family of sets. Let us denote this family by $A_{\rm opt}$.

$1^\circ$. Let us first show that this family $A_{\rm opt}$ is indeed $G$-invariant, i.e., that $g(A_{\rm opt}) = A_{\rm opt}$ for every transformation $g \in G$.

Indeed, let $g \in G$. From the optimality of $A_{\rm opt}$, we conclude that for every $B \in \mathcal{A}$, $g^{-1}(B) \preceq A_{\rm opt}$. From the $G$-invariance of the optimality criterion, we can now conclude that $B \preceq g(A_{\rm opt})$. This is true for all $B \in \mathcal{A}$ and therefore, the family $g(A_{\rm opt})$ is optimal. But since the criterion is final, there is only one optimal family; hence, $g(A_{\rm opt}) = A_{\rm opt}$. So, $A_{\rm opt}$ is indeed invariant.

$2^\circ$. Let us now show an arbitrary set $X_0$ from the optimal family $A_{\rm opt}$ consists of orbits of $\geq (d-r)$-dimensional subgroups of the group $G$.

Indeed, the fact that $A_{\rm opt}$ is $G$-invariant means, in particular, that for every $g \in G$, the set $g(X_0)$ also belongs to $A_{\rm opt}$. Thus, we have a (smooth) mapping $g \to g(X_0)$ from the $d$-dimensional manifold $G$ into the $\leq r$-dimensional set $G(X_0) = \{g(X_0) \,|\, g \in G\} \subseteq A_{\rm opt}$. In the following, we will denote this mapping by $g_0$.

Since $r < d$, this mapping cannot be 1-1, i.e., for some sets $X = g'(X_0) \in G(X_0)$, the pre-image $g_0^{-1}(X) = \{g \,|\, g(X_0) = g'(X_0)\}$ consists of one than one point. By definition of $g(X)$, we can conclude that $g(X_0) = g'(X_0)$ iff $(g')^{-1}g(X_0) = X_0$. Thus, this pre-image is equal to $\{g \,|\, (g')^{-1}g(X_0) = X_0\}$. If we denote $(g')^{-1}g$ by $\tilde{g}$, we conclude that $g = g'\tilde{g}$ and that the pre-image $g_0^{-1}(X) = g_0^{-1}(g'(X_0))$ is equal to $\{g'\tilde{g} \,|\, \tilde{g}(X_0) = X_0\}$, i.e., to the result of applying $g'$ to $\{\tilde{g} \,|\, \tilde{g}(X_0) = X_0\} = g_0^{-1}(X_0)$. Thus, each pre-image ($g_0^{-1}(X) = g_0^{-1}(g'(X_0))$) can be obtained from one of these pre-images (namely, from $g_0^{-1}(X_0)$) by a smooth invertible transformation $g'$. Thus, all pre-images have the same dimension $D$.

We thus have a *stratification* (fiber bundle) of a $d$-dimensional manifold $G$ into $D$-dimensional strata, with the dimension $D_f$ of the factor-space being $\le r$. Thus, $d = D + D_f$, and from $D_f \le r$, we conclude that $D = d - D_f \ge n - r$.

So, for every set $X_0 \in A_{\text{opt}}$, we have a $D \ge (n-r)$-dimensional subset $G_0 \subseteq G$ that leaves $X_0$ invariant (i.e., for which $g(X_0) = X_0$ for all $g \in G_0$). It is easy to check that if $g, g' \in G_0$, then $gg' \in G_0$ and $g^{-1} \in G_0$, i.e., that $G_0$ is a *subgroup* of the group $G$. From the definition of $G_0$ as $\{g \mid g(X_0) = X_0\}$ and the fact that $g(X_0)$ is defined by a smooth transformation, we conclude that $G_0$ is a smooth sub-manifold of $G$, i.e., a $\ge (n-r)$-dimensional subgroup of $G$.

To complete our proof, we must show that the set $X_0$ is a union of orbits of the group $G_0$. Indeed, the fact that $g(X_0) = X_0$ means that for every $x \in X_0$, and for every $g \in G_0$, the element $g(x)$ also belongs to $X_0$. Thus, for every element $x$ of the set $X_0$, its entire orbit $\{g(x) \mid g \in G_0\}$ is contained in $X_0$. Thus, $X_0$ is indeed the union of orbits of $G_0$. The proposition is proven.

## Acknowledgements

This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721 and by Grant 1 T36 GM078000-01 from the National Institutes of Health.

## References

1. Branden, C. I.; Tooze, J. *Introduction to Protein Structure*; Garland Publ., New York, 1999.
2. Finkelstein, A.; Kosheleva, O.; Kreinovich, V. Astrogeometry, error estimation, and other applications of set-valued analysis. *ACM SIGNUM Newsletter* **1996**, *31*(4), 3-25.
3. Finkelstein, A.; Kosheleva, O.; Kreinovich, V. Astrogeometry: towards mathematical foundations. *International Journal of Theoretical Physics* **1997**, *36*(4), 1009-1020.
4. Finkelstein, A.; Kosheleva, O.; Kreinovich, V. Astrogeometry: geometry explains shapes of celestial bodies. *Geombinatorics* **1997**, *VI*(4), 125-139.
5. Gromov, M. Crystals, proteins and isoperimetry. *Bulletin of the American Mathematical Society* **2011**, *48*(2), 229-257.
6. Lesk, A. M. *Introduction to Protein Science: Architecture, Function, and Genomics*; Oxford University Press, New York, 2010.
7. Li, S.; Ogura, Y.; Kreinovich, V. *Limit Theorems and Applications of Set Valued and Fuzzy Valued Random Variables*, Kluwer Academic Publishers, Dordrecht, 2002.
8. Novotny, J.; Bruccoleri, R. E.; Newell, J. Twisted hyperboloid (Strophoid) as a model of beta-barrels in proteins. *J. Mol. Biol.* **1984**, *177*, 567-573.
9. Stec, B.; Kreinovich, V. Geometry of protein structures. I. Why hyperbolic surfaces are a good approximation for beta-sheets. *Geombinatorics* **2005**, *15*(1), 18-27.