Submitted to *Symmetry*. Pages 1 - 11.

OPEN ACCESS

Article

Towards Symmetry-Based Explanation of (Approximate) Shapes of Alpha-Helices and Beta-Sheets (and Beta-Barrels) in Protein Structure

Jaime Nava and Vladik Kreinovich*

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

* Author to whom correspondence should be addressed; vladik@utep.edu, Tel. +1-915-747-6951, Fax +1-915-747-5030.

Version January 5, 2012 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 sometimes 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, correspondingly, cylindrical spirals and planes (and cylinders, for beta-barrels). In this paper, following the ideas pioneered by a renowned mathematician M. Gromov, we use natural symmetries to show that, under reasonable assumptions, these geometric shapes are indeed the best approximating families for secondary structures.

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

1. Introduction

Alpha-helices and bet-sheets: brief reminder. 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 *line-like* feature defines a secondary structure called an *alpha-helix*. 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., [1,9].

Shapes of alpha-helices and bet-sheets: first approximation. The actual shapes of the alpha-helices and beta-sheets can be complicated. In the first approximation, alpha-helices are usually approximated by *cylindrical spirals* (also known as *circular helices* or (cylindrical) *coils*), i.e., curves which, in an appropriate coordinate system, have the form $x = a \cdot \cos(\omega \cdot t)$, $y = a \cdot \sin(\omega \cdot t)$, and $c = b \cdot t$. Similarly, in the first approximation, beta-sheets are usually approximated as *planes*. These are the shapes that we will try to explain in this paper.

What we do in this paper: our main result. In this paper, following the ideas of a renowned mathematician M. Gromov [8], we use symmetries to show that under reasonable assumptions, the empirically observed shapes of cylindrical spirals and planes are indeed the best families of simple approximating sets.

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

Auxiliary result: we also explain the (approximate) shape of beta-barrels. The actual shape of an alpha-helix or of a beta-sheet is somewhat different from these first-approximation shapes. In [12], we showed that symmetries can explain some resulting shapes of beta-sheets. In this paper, we will add, to the basic approximate shapes of a circular helix and a planes, one more shape. This shape is observed when, due to tertiary structure effects, a beta-sheet "folds" on itself, becoming what is called a *beta-barrel*. In the first approximation, beta-barrels are usually approximated by cylinders. So, in this paper, we will also explain cylinders.

We hope that similar symmetry ideas can be used to describe other related shapes. For example, it would be nice to see if a torus shape – when a cylinder folds on itself – can also be explained by symmetry ideas.

Possible future work: need for explaining shapes of combinations of alpha-helices and beta-sheets. A protein usually consists of several alpha-helices and beta-sheets. In some cases, these combinations of basic secondary structure elements have their own interesting shapes: e.g., coils (alpha-helices) sometimes form a *coiled coil.* In this paper, we use symmetries to describe the basic geometric shape of secondary structure elements; we hope that similar symmetry ideas can be used to describe the shape of their combinations as well.

2. Symmetry Approach in Physics: Brief Reminder

Symmetries are actively used in physics. In our use of symmetries, we have been motivated by the successes of using symmetries in physics; see, e.g., [2]. So, in order to explain our approach, let us first briefly recall how symmetries are used in physics.

Symmetries in physics: main idea. In physics, we usually know the differential equations that describe the system's dynamics. Once we know the initial conditions, we can then solve these equations and obtain the state of the system at any given moment of time.

It turns out that in many physical situations, there is no need to actually solve the corresponding complex system of differential equations: the same results can be obtained much faster if we take into account that the system has certain *symmetries* (i.e., transformations under which this system does not change).

Symmetries in physics: examples. Let us give two examples of the use of symmetries in physics:

- a simpler example in which we will be able to perform all the computations, and
- a more complex example in which we will skip all the computations and proofs – but which will be useful for our analysis of the shape of proteins.

First example: pendulum. As the first simple example, let us consider the problem of finding how the period T of a pendulum depends on its length L and on the free fall acceleration g on the corresponding planet. We will denote the desired dependence by $T = f(L, g)$. This dependence was originally found by using Newton's equations. We will show that (modulo a constant) the same dependence can be obtained without using any differential equations, only by taking the corresponding symmetries into account.

What are the natural symmetries here? To describe a numerical value of the length, we need to select a unit of length. In this problem, there is no fixed length, so it makes sense to assume that the physics does not change if we simply change the unit of length. If we change a unit of length to a one λ times smaller, we get new numerical value $L' = \lambda \cdot L$; e.g., 1.7 m = 170 cm.

Similarly, if we change a unit of time to a one which is μ times smaller, we get a new numerical value for the period $T' = \mu \cdot T$. Under these transformations, the numerical value of the acceleration changes as $g \to g' = \lambda \cdot \mu^{-2} \cdot g$.

Since the physics does not change by simply changing the units, it makes sense to require that the dependence $T = f(L, g)$ also does not change if we simply change the units, i.e., that $T = f(L, g)$ implies $T' = f(L', g')$. Substituting the above expressions for T', L', and g' into this formula, we conclude that $f(\lambda \cdot L, \lambda \cdot \mu^{-2} \cdot g) = \mu \cdot f(L, g)$. From this formula, we can find the explicit expression for the desired function $f(L, g)$. Indeed, let us select λ and μ for which $\lambda \cdot L = 1$ and $\lambda \cdot \mu^{-2} \cdot g = 1$. Thus, we take $\lambda = L^{-1}$ and $\mu = \sqrt{\lambda \cdot g} = \sqrt{g/L}$. For these values λ and μ, the above formula takes the form $f(1, 1) = \mu \cdot f(L, g) = \sqrt{g/L} \cdot f(L, g)$. Thus, $f(L, g) = \text{const} \cdot \sqrt{L/g}$ (for the constant $f(1, 1)$). This is exactly the same formula that we obtain from Newton's equations.

What is the advantage of using symmetries? At first glance, the above derivation of the pendulum formula is somewhat useless: we did not invent any new mathematics, the above mathematics is very simple, and we did not come up with any new physical conclusion – the formula for the period of the pendulum is well known. Yes, we got a slightly simpler derivation, but once a result is proven, getting a new shorter proof is not very interesting. So what is new in this derivation?

What is new is that we derived the above without using any specific differential equations – we only the fact that these equations do not have any fixes unit of length or fixed unit of time. Thus, the same

formula is true not only for Newton's equations, but also for *any* alternative theory – as long as this alternative theory has the same symmetries.

Another subtle consequence of our result is related to the fact that physical theories need to be experimentally confirmed. Usually, when a formula obtained from a theory turned out to be experimentally true, this is a strong argument for confirming that the original theory is true. One may similarly think that if the pendulum formula is experimentally confirmed, this is a strong argument for confirming that Newton's mechanics is true. However, the fact that we do not need the whole theory to derive the pendulum formula – we only need symmetries – shows that:

- if we have an experimental confirmation of the pendulum formula,
- this does not necessarily mean that we have confirmed Newton's equations – all we confirmed are the symmetries.

General comment about physical problems and fundamental physical equations. The fact that we could derive this formula so easily – shows that maybe in more complex situations, when solving the corresponding differential equation is not as easy, we would still be able to find an explicit solution by using appropriate symmetries. This is indeed the case in many complex problems; see, e.g., [2].

Moreover, in many situations, even equations themselves can be derived from the symmetries. This is true for most equations of fundamental physics: Maxwell's equations of electrodynamics, Einstein's General Relativity equations for describing the gravitation field, Schrödinger's equations of quantum mechanics, etc.; see, e.g., [6,7].

As a result, in modern physics, often, new theories are formulated not in terms of differential equations, but in term of symmetries. This started with quarks whose theory was first introduced by M. Gell-Mann by postulating appropriate symmetries.

Second example: shapes of celestial objects. Another example where symmetries are helpful is the description of observed geometric shapes of celestial bodies. Many galaxies have the shape of planar logarithmic spirals; other clusters, galaxies, galaxy clusters have the shapes of the cones, conic spirals, cylindrical spirals, straight lines, spheres, etc. For several centuries, physicists have been interested in explaining these shapes. For example, there exist several dozen different physical theories that explain the observed logarithmic spiral shape of many galaxies. These theories differ in their physics, in the resulting differential equations, but they all lead to exactly the same shape – of the logarithmic spiral.

It turns out that there is a good explanation for this phenomenon – all observed shapes can be deduced from the corresponding symmetries; see, e.g., [3–5,10]. Here, possible symmetries include shifts, rotations, and "scaling" (dilation) $x_i \to \lambda \cdot x_i$.

The fact that the shapes can be derived from symmetry shows that the observation of these shapes does not confirm one of the alternative theories – it only confirms that all these theories are invariant under shift, rotation, and dilation. This derivation also shows that even if the actual physical explanation for the shape of the galaxies turns out to be different from any of the current competing theories, we should not expect any new shapes – as long as we assume that the physics is invariant with respect to the above basic geometric symmetries.

3. From Physics to Analyzing Shapes of Proteins: Towards the Formulation of the Problem

Reasonable symmetries. It is reasonable to assume that the underlying chemical and physical laws do not change under shifts and rotations. Thus, as a group of symmetries, we take the group of all "solid motions", i.e., of all transformations which are composed of shifts and rotations.

Comment. In the classification of shapes of celestial bodies, we also considered dilations. Dilations make sense in astrophysics and cosmology. Indeed, in forming celestial shapes of large-scale objects, the main role is played by long-distance interactions like gravity and electromagnetic forces, and the formulas describing these long-distance interactions are dilation-invariant. In constant, on the molecular level – that corresponds to the shapes of the proteins – short-distance interactions are also important, and these interactions are not necessarily dilation-invariant.

Thus, in our analysis of protein shapes, we only consider shifts and rotations.

Reasonable shapes. In chemistry, different shapes are possible. For example, *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 potentially *unbounded* shapes. Specifically, we want to describe *connected* components of these shapes.

Reasonable families of shapes. We do not want to just find one single shape, we want to find *families* of shapes that approximate the actual shapes of proteins. These families contain several parameters, so that by selecting values of all these parameters, we get a shape.

The more parameters we allow, the larger the variety of the resulting shape and therefore, the better the resulting shape can match the observed protein shape.

We are interested in the shapes that describe the secondary structure, i.e., the first (crude) approximation to the actual shape. Because of this, we do not need too many parameters, we should restrict ourselves to families with a few parameters.

We want to select the best approximating family. In principle, we can have many different approximating families. Out of all these families, we want to select a one which is the *best* in some reasonable sense – e.g., the one that, on average, provides the most accurate approximation to the actual shape, or the one which is the fastest to compute, etc.

What does the "best" mean? There are many possible criteria for selecting the "best" family. It is not easy even to enumerate all of them – while our objective is to find the families which are the best according to each of these criteria. To overcome this difficulty, we therefore formulate a *general* description of the optimality criteria and provide a general description of all the families which are optimal with respect to different criteria.

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 $A_{\rm opt}$, i.e., the unique family for which $\forall B\,(A_{\rm opt} \succeq B)$. 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 ($A \succeq_{\rm new} B$ if either A gives a better approximation, or if $A \sim_{\rm old} B$ 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 also reasonable to require that the relation $A \succeq B$ should be invariant relative to natural geometric symmetries, i.e., that this relation is shift- and rotation-invariant.

At fist glance, these requirements sounds reasonable but somewhat weak. We will show, however, that they are sufficient to actually find the optimal families of shapes – and that the resulting optimal shapes are indeed the above-mentioned observed secondary-structure shapes of protein components.

4. Definitions and the Main 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, 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 $\{g(X) \,|\, X \in A\}$.*

In our problem, the set M is the 3-D space $\mathbb{R}^3$.

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.*

In our problem, the group G is the group generated by all shifts and rotations. in the 3-D space, we need three parameters to describe a general shift, and three parameters to describe a general rotation; thus, the group G is 6-dimensional – in the sense that we need six parameters to describe an individual element of this group.

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* π*) 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, Π) *for which* B *consists of all the sets* $s(v)$ *with* v *from some* r*-dimensional sub-manifold* $W \subseteq V$.

In our example, we consider families of unbounded connected sets.

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* $B \neq A$.

Lemma. *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_{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, all the proofs are placed in the following Proofs section.

Theorem. *Let G be a 6-dimensional group generated by all shifts and rotations in the 3-D space $\mathbb{R}^3$, and let $\preceq$ be a G-invariant and final optimality criterion on the class $\mathcal{A}$ of all r-parametric families of unbounded sets from $\mathbb{R}^3$, $r < 6$. Then each set X from the optimal family is a union of cylindrical spirals, planes, and cylinders.*

Conclusion. 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., [11]. It is worth mentioning that the empirical fact that quadratic models provide the best second approximation can also be theoretical explained by using symmetries [12].

5. Proofs

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

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

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

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

Indeed, the fact that A_{opt} is G-invariant means, in particular, that for every $g \in G$, the set $g(X_0)$ also belongs to A_{opt}. Thus, we have a (smooth) mapping $g \to g(X_0)$ from the d-dimensional manifold G into the $\leq r$-dimensional set $G(X_0) = \{g(X_0) \,|\, g \in G\} \subseteq A_{\text{opt}}$. 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 $\leq r$. Thus, $d = D + D_f$, and from $D_f \leq r$, we conclude that $D = d - D_f \geq n - r$.

So, for every set $X_0 \in A_{\rm opt}$, we have a $D \geq (n-r)$-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 $\{g \,|\, g(X_0) = X_0\}$ 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 $\geq (n-r)$-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 $\{g(x) \,|\, g \in G_0\}$ is contained in X_0. Thus, X_0 is indeed the union of orbits of G_0. The lemma is proven.

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

Since we are interested in connected components, we should consider only connected *continuous* subgroups $G_0 \subseteq G$, since such 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 κ_i, they may only shift s (to $s + s_0$). This means that the values of κ_i are constant. Taking constant κ_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*.

Since we are only interested in unbounded shapes, 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.

The theorem is proven.

6. Symmetry-Related Speculations on Possible Physical Origin of the Observed Shapes

We have provided a somewhat mathematical explanation for the observed shapes. Our theorem explains the shapes, but not how a protein acquires these shapes.

A possible (rather speculative) explanation can be obtained along the lines of a similar symmetry-based explanation for the celestial shapes; see [3–5,10].

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.

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. The authors are thankful to the anonymous referees for valuable suggestions.

References

1. Branden, C. I.; Tooze, J. *Introduction to Protein Structure*; Garland Publ., New York, 1999.
2. Feynman, R. P.; Leighton, R. B.; Sands, M. *Feynman Lectures on Physics*, Addison- Wesley, Boston, Massachusetts, 2005.
3. Finkelstein, A.; Kosheleva, O.; Kreinovich, V. Astrogeometry, error estimation, and other applications of set-valued analysis. *ACM SIGNUM Newsletter* **1996**, *31*(4), 3-25.
4. Finkelstein, A.; Kosheleva, O.; Kreinovich, V. Astrogeometry: towards mathematical foundations. *International Journal of Theoretical Physics* **1997**, *36*(4), 1009-1020.
5. Finkelstein, A.; Kosheleva, O.; Kreinovich, V. Astrogeometry: geometry explains shapes of celestial bodies. *Geombinatorics* **1997**, *VI*(4), 125-139.
6. Finkelstein, A. M.; Kreinovich, V. Derivation of Einstein's, Brans-Dicke and other equations from group considerations. In: Choque-Bruhat, Y.; Karade, T. M. (eds) *On Relativity Theory. Proceedings of the Sir Arthur Eddington Centenary Symposium, Nagpur India 1984*, World Scientific, Singapore, **1985** *2*, 138-146.
7. Finkelstein, A. M.; Kreinovich, V.; Zapatrin, R. R. Fundamental physical equations uniquely determined by their symmetry groups. *Springer Lecture Notes in Mathematics* **1986**, *1214*, 159-170.
8. Gromov, M. Crystals, proteins and isoperimetry. *Bulletin of the American Mathematical Society* **2011**, *48*(2), 229-257.
9. Lesk, A. M. *Introduction to Protein Science: Architecture, Function, and Genomics*; Oxford University Press, New York, 2010.
10. Li, S.; Ogura, Y.; Kreinovich, V. *Limit Theorems and Applications of Set Valued and Fuzzy Valued Random Variables*, Kluwer Academic Publishers, Dordrecht, 2002.
11. 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.
12. 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.