Mini Review

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2023.50.007914

Protein Folding – Quantification of Protein Complexity, Robustness and Amino Acid Participation

Marczyk J[1]*, Stadler P[2] and Treguerres JAF[3]

[1]*Ontonix S.r.l. Como, Italy*

[2]*Biotechnology Consultant, Germany*

[3]*Faculty of Medicine, Complutense University of Madrid, Spain*

***Corresponding author:** Marczyk J, Ontonix S.r.l. Como, Italy*

ARTICLE INFO

Received: April 24, 2023

Published: May 03, 2023

Citation: Marczyk J, Stadler P and Treguerres JAF. Protein Folding – Quantification of Protein Complexity, Robustness and Amino Acid Participation. Biomed J Sci & Tech Res 50(2)-2023. BJSTR. MS.ID.007914.

ABSTRACT

The QCT (Quantitative Complexity Theory) algorithm has been applied to the analysis of the folding process of a protein composed of 435 atoms, monitoring its complexity at each step thereof. The folding has been simulated using Molecular Dynamics Simulation. The analysis has revealed that, while in the native state, the protein's configuration minimizes its energy, its complexity reaches a maximum. This result is interesting in that, according to the QCT, complexity is information that is encoded in structure. It is conjectured that the native state of a protein is a minimum energy-maximum information state. Moreover, QCT allows us to determine the footprint of each constituent amino acid in the dynamics, information content and robustness of a protein's structure. The application of QCT on proteins generates data and information about structure, complexity, special arrangements, etc., of proteins. The knowledge about the biological functions of such proteins derived from the above – which is crucial e.g., for designing new drugs - will have to be generated in collaboration with specialists from pharmaceutical R&D.

Introduction and Background

Since its introduction in 2005, the Quantitative Complexity Theory (QCT) provides quantitative and holistic information on the state of generic multi-functional dynamic systems [1]. Its applicability and advantages in medicine have been demonstrated in predicting clinical events in cardiac resynchronisation therapy [2,3], or of the outcome of the Head Up Tilt Test [4], where complexity profiling has provided a detailed assessment of individual hemodynamic patterns of syncope. Complexity has been shown to be a sensitive marker of a cardiovascular hemodynamic response to orthostatic stress and vasodilator administration, and its increase has preceded changes in standard cardiovascular parameters [5]. A recent application of the QCT is in the field of Molecular Dynamics [6]. By processing spatial trajectories of atoms in a molecule, the QCT allows to measure the complexity of a molecule, its robustness, and provides a detailed breakdown of how information is encoded in the molecule in the form of a graphical representation, known as a Complexity Map. Furthermore, the application of the QCT to the results of Molecular Dynamics Simulations allows us to establish a molecule ranking mechanism in terms of complexity, robustness and, potentially, use them to predict certain physical-chemical properties.

Proteins are modular constructs. They are formed by chaining amino acid molecules, with the smallest amino acid (glycine, C_2H_3NO) having 7 atoms and the largest (pyrrolysine, $C_{12}H_{19}N_3O_2$) having 36. Depending on how many amino acids there are in a protein (ranging from a few dozen to several thousand), the number of atoms varies immensely. The smallest known protein is TRP-cage, with only 20 amino acids and 154 atoms. Understanding and simulating the protein folding process has been an important challenge for computational biology since the late 1960s. The present paper describes the application of the QCT to the analysis of the folding process of a protein composed of 435 atoms. The process of folding has been simulated in an aqueous environment using Molecular Dynamics Simulation from the unfolded to the native state, producing the x, y and z coordinates of each atom over time.

The Quantitative Complexity Theory

Complexity is a new descriptor of a system. The complexity of a system described by a vector {x} of N components is defined as a scalar function of Structure and Entropy as follows: C = f (S ○E), where S represents an N × N adjacency matrix, E is an N × N entropy matrix, "○" is the Hadamard matrix product operator and f is a matrix norm operator. Since complexity is a function of entropy, and given that S has no units, its units are bits. The above equation represents a formal definition of complexity, and it is not used in its computation. The adjacency matrix is determined via a propriety multi-dimensional algorithm which determines if entry Sij is 0 or 1. This establishes the structure of the system in question. The intensity of relationships between the components of {x}, the so-called generalized correlation, is computed based on concepts of entropy and Shannon's Information Theory, [6]. This approach has been chosen because it avoids the drawbacks of conventional techniques whereby one analyses data via regression models, cluster analysis or other methods. The huge advantage of this "model-free" approach is that it is independent of numerical conditioning of the data and its ability to identify the existence of structures where conventional methods fail. Once the entropy matrix and the adjacency matrix have been obtained, one may compute the complexity of a given system as the following matrix norm: C = || S ○ E ||. A fundamental property of the above measure of complexity is that it is bounded. The upper bound, called critical complexity, C_U, as well as the lower bound, C_L, are also computed based on proprietary algorithms. The robustness of a system, R, may be computed as a function of the ratio (C - C_U) / (C_U - C_L) and ranges from 0 to 100%. In proximity of the lower complexity bound, a given system functions in a deterministic structure-dominated fashion. In proximity of critical complexity system functioning is chaotic and relationships between the various entries of {x} are fuzzy and therefore characterized by very low generalized correlations. This means that the structure, S, is feeble and therefore has low robustness.

From Molecular Dynamics to Protein Complexity

Molecular Dynamics simulation, first developed in the late 70s, has advanced from simulating structures several hundreds of atoms to systems with biological relevance, including entire proteins in solution with explicit solvent representations, membrane embedded proteins, or large macromolecular complexes like nucleosomes or ribosomes. Simulation of systems having ~50,000–100,000 atoms are now routine, and simulations of more than 500,000 atoms are common when appropriate computer facilities are available. This remarkable improvement is in large part a consequence of the use of high-performance computing (HPC), and the simplicity of the basic MD algorithm (see figure below) (Figure 1). An initial model of the system is obtained from either experimental structures or comparative modelling data. Once the system is built, forces acting on every atom are obtained by deriving equations, the force-fields, where potential energy is deduced from the molecular structure. Once the forces acting on individual atoms are obtained, classical Newton's law of motion is used to calculate accelerations and velocities and to update the atom positions. As integration of movement is done numerically, to avoid instability, a time step shorter than the fastest movements in the molecule should be used. This ranks normally between 1 and 2 fs for atomistic simulations.

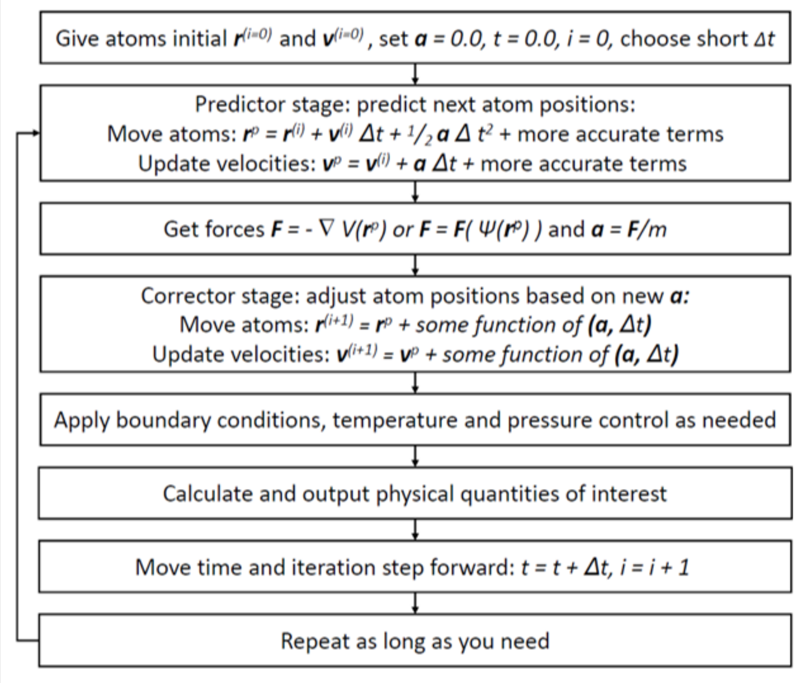
Give atoms initial $\boldsymbol{r}^{(i=0)}$ and $\boldsymbol{v}^{(i=0)}$, set $\boldsymbol{a} = 0.0$, $t = 0.0$, $i = 0$, choose short Δt
Predictor stage: predict next atom positions:
Move atoms: $\boldsymbol{r}^p = \boldsymbol{r}^{(i)} + \boldsymbol{v}^{(i)} \Delta t + {}^1/_2 \boldsymbol{a} \Delta t^2$ + more accurate terms
Update velocities: $\boldsymbol{v}^p = \boldsymbol{v}^{(i)} + \boldsymbol{a} \Delta t$ + more accurate terms
Get forces $\boldsymbol{F} = - \nabla V(\boldsymbol{r}^p)$ or $\boldsymbol{F} = \boldsymbol{F}(\Psi(\boldsymbol{r}^p))$ and $\boldsymbol{a} = \boldsymbol{F}/m$
Corrector stage: adjust atom positions based on new $\boldsymbol{a}$:
Move atoms: $\boldsymbol{r}^{(i+1)} = \boldsymbol{r}^p$ + some function of $(\boldsymbol{a}, \Delta t)$
Update velocities: $\boldsymbol{v}^{(i+1)} = \boldsymbol{v}^p$ + some function of $(\boldsymbol{a}, \Delta t)$
Apply boundary conditions, temperature and pressure control as needed
Calculate and output physical quantities of interest
Move time and iteration step forward: $t = t + \Delta t$, $i = i + 1$
Repeat as long as you need

Figure 1: Simplified scheme of the Molecular Dynamics Simulation algorithm.

DOI: 10.26717/BJSTR.2023.50.007914

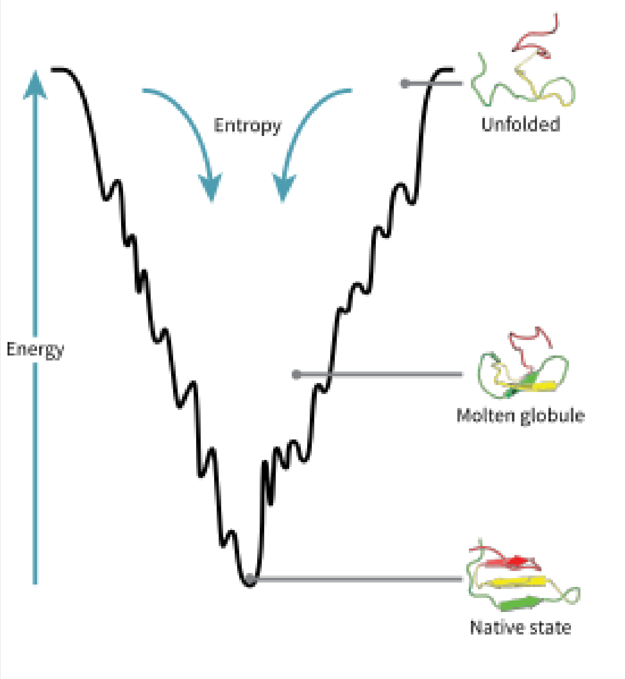

Figure 2: A protein's energy well, showing the unfolded and native states.

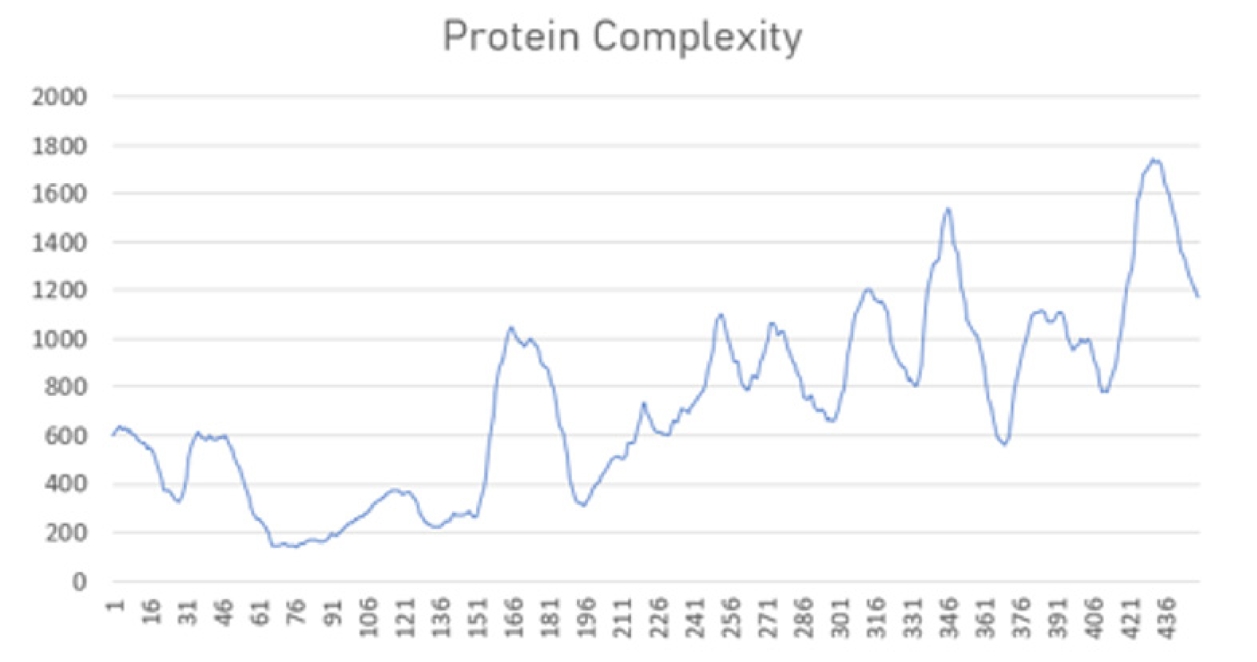

Figure 3: Evolution of protein's complexity over time. As the protein approaches its native state its complexity increases.

Proteins tend towards their native state, which occupies a (local) minimum energy position in the protein's energy landscape, see figure below (Figure 2): The QCT algorithm has been applied to a protein comprised of 435 atoms. Molecular Dynamics simulation has been utilized to determine the position of each of the atoms as the protein transitioned from its unfolded to the native state. The data has been provided by the CINECA Supercomputer Centre in Bologna, Italy (https://www.cineca.it/). The simulation output has been analyzed with the QCT algorithm using a moving window. The evolution of complexity of the protein is illustrated in the figure below (Figure 3). It may be observed that complexity tends towards a maximum. While a protein tends toward a minimum-energy configuration, at the same time it maximizes its own complexity. This makes sense if we recall the definition of complexity: complexity is structured information, or information that emanates from structure. In multidimensional systems each dimension carries information independently of other dimensions. However, the additional information that is derived from the interdependencies between the different dimensions is often significantly larger than that of the single dimensions. The shape (structure) of a protein encodes information. For given ensembles of amino acids, which will ultimately fold into a protein, two things happen: the protein's energy is minimized and, at the same time, the amount of information that it can encode with those amino acids is maximized. With all likelihood – this is only a conjecture – the final state of a protein is a compromise (combination) of these two characteristics: minimum energy and maximum information.

The importance of structure cannot be overstated. Structure drives functionality and in the context of the QCT, structure is defined by the exchange (flow) of information between the various dimensions of a given system. Systems with more structure can perform more functions. Not surprisingly, in our biosphere there is a drive towards states of higher complexity (higher functionality) and higher robustness (fitness). The QCT not only establishes this structure, S, it also measures the amount of information it carries, C, as well as its robustness, R. The figure below illustrates a simple example of a hypothetical protein, composed of 6 amino acids (for a total of 26 atoms), its Complexity Map and, the Complexity Profile, which ranks the information footprint of each amino acid in the entire protein. In other words, the Complexity Profile indicates which amino acids are the drivers of the overall dynamic properties of the protein. In the example below, amino acid 1 carries almost 35% of the complexity of the protein, amino acid 2 close to 26% and amino acid 3 just under 25%. In essence, these three amino acids drive over 85% of the proteins structure, its dynamics, as well as the protein's robustness and stability (Figure 4).

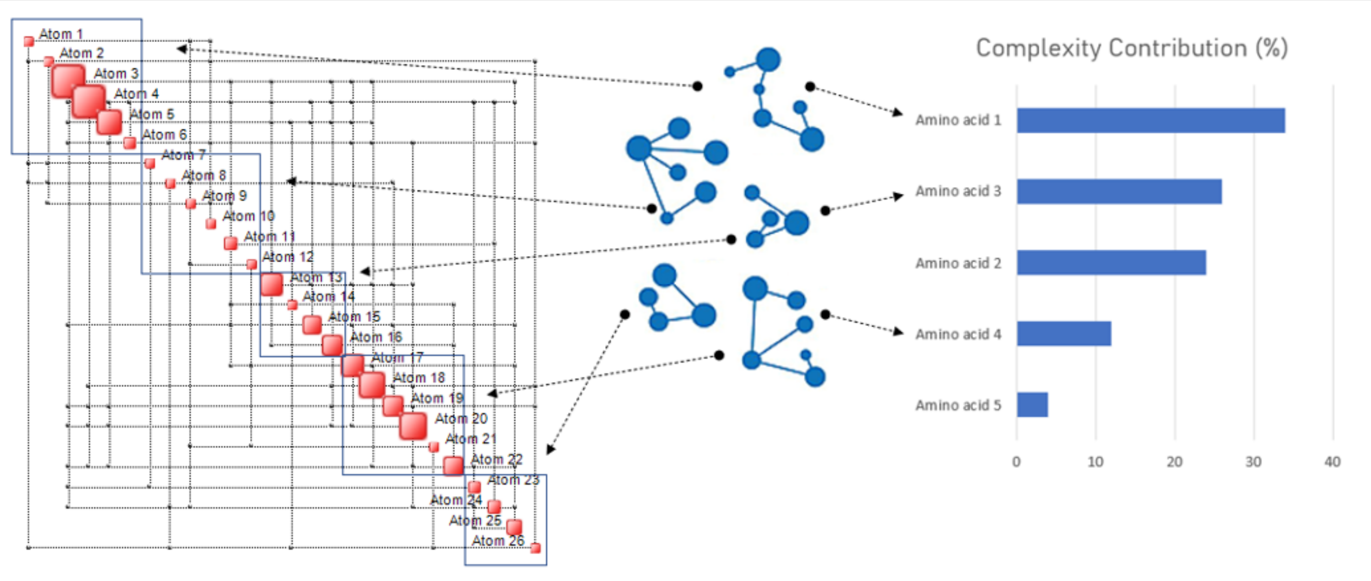

Figure 4: Example of a hypothetical protein's Complexity Map. The protein's Complexity Profile on the right-hand side provides a ranking of the degree of participation of each amino acid in the protein's complexity and its dynamics.

The evolution of the Complexity Map of the 435-atom protein is shown below. The number of nodes on the diagonal is 435 x 3 = 1305, which corresponds to the number of degrees of freedom (Figure 5). The map on the right-hand side, corresponding to the native state of the protein, is shown below. The entries of the map's diagonal are the x, y, z coordinates of each atom and off-diagonal dots represent the presence of dynamic coupling between degrees of freedom (Figure 6). While the above results need to be verified by analyzing other proteins, it appears that a minimum energy-maximum complexity (information) state is an attractor which results from the process of folding. The more information a protein encodes, the more functions it can potentially perform. The protein's Complexity Landscape is illustrated below. The landscape reflects the evolution of the protein's Complexity Profile over time. One may notice how in the initial, unfolded configuration, there are certain amino acids that dominate the dynamics of the folding process. However, the footprint (participation factor) of each amino acid in the protein's native state is distributed more evenly (Figure 7).

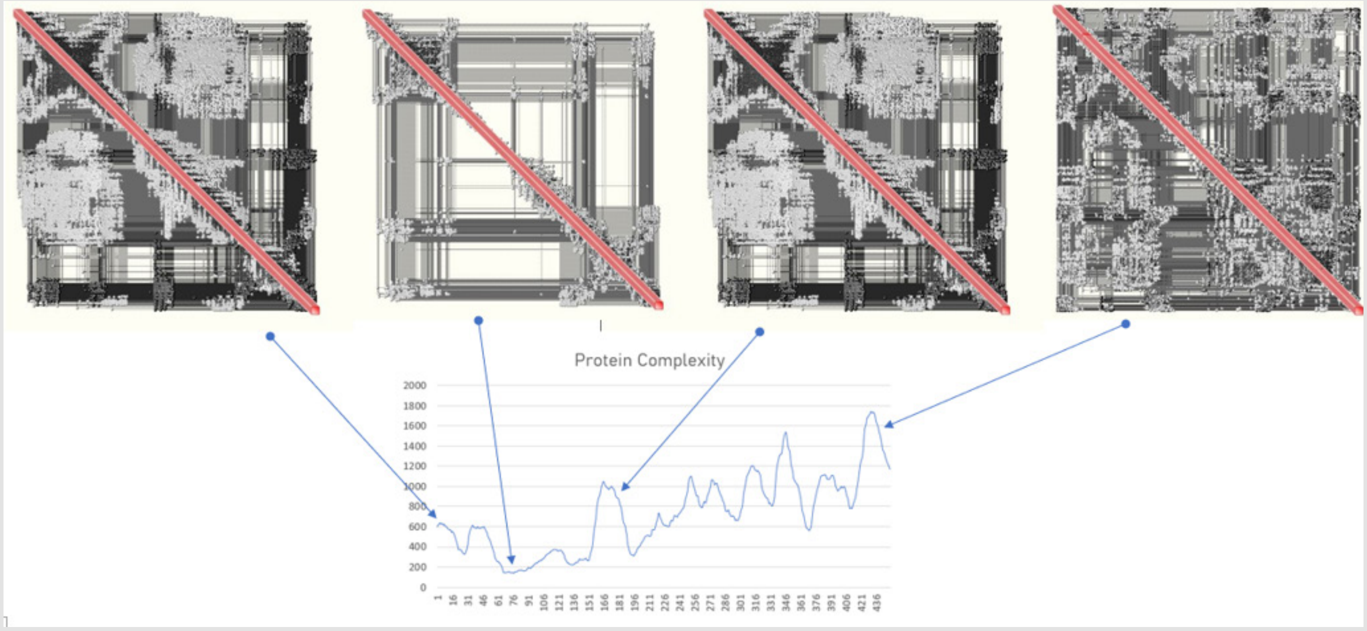

Figure 5: Evolution of the structure of the protein's Complexity Map. The map provides a graphical representation of how information is encoded in the protein. QCT analysis of the 435-atom protein requires just over 20 hours on an i9 CPU, and around 10 minutes on a 128-core system.

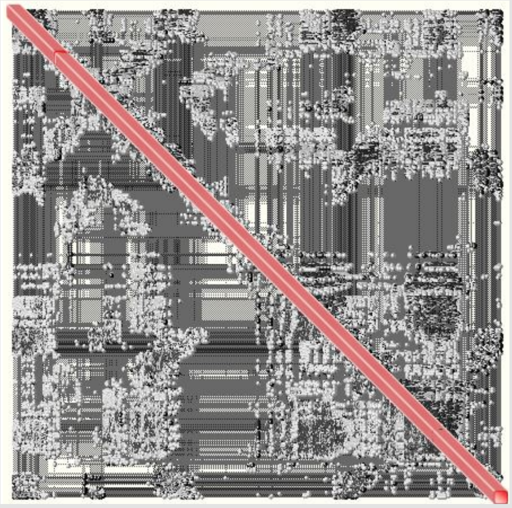

Figure 6: Structure of the protein's Complexity Map in proximity of its native state. The robustness of the protein's structure is R=60% and it may be linked to the protein's life span.

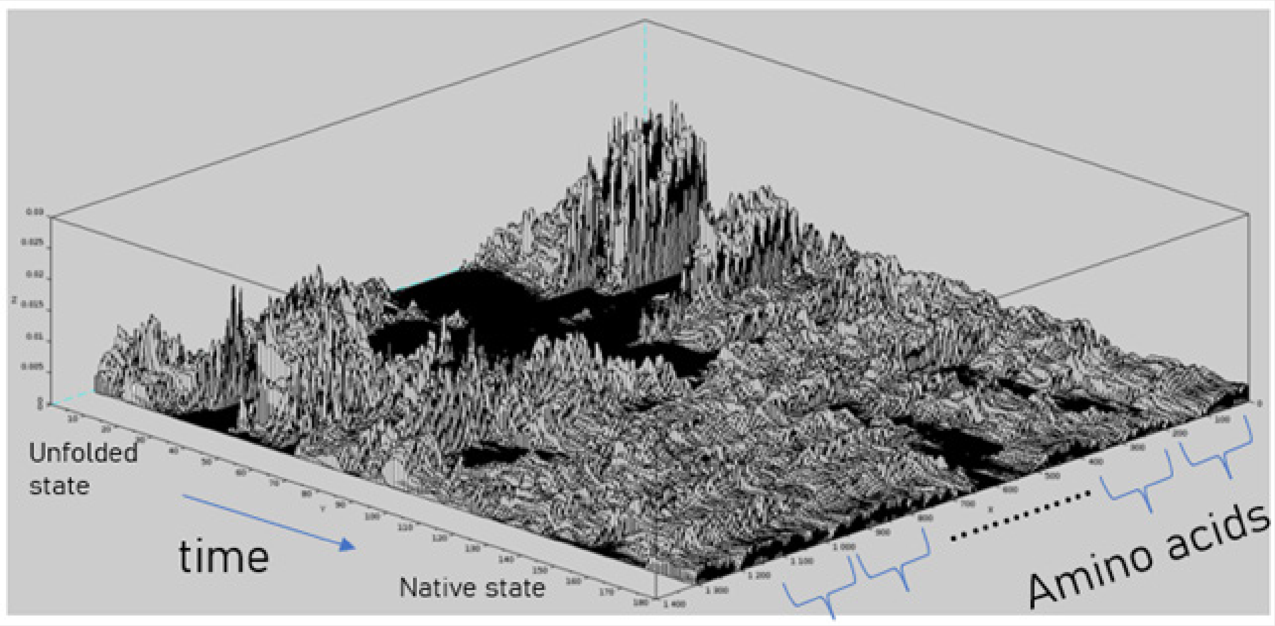

Figure 7: Protein's Complexity Landscape. In the initial state there were a few 'dominant' atoms (amino acids) while in the native state, the amino acid participation factors are distributed more uniformly.

Future Work

In order to verify that in its native state a protein assumes a condition of maximum complexity – the encoded information is maximum – the processing of a larger number of proteins will be required. Moreover, the goal is to analyse the complexity of selected proteins while they perform functions such as chemical messengers or enzymes. In these contexts, the QCT algorithm in conjunction with Molecular Dynamics Simulation is able to furnish the following information:

1. Measure the complexity of a protein. This provides insights into the way information is encoded and distributed spatially. This can be performed at amino acid or atomic level.
2. Measure the robustness of the protein's structure, indicating concentrations of fragility. This can be performed at amino acid or atomic level. A protein's life span and stability may be linked to its robustness, and it is possible to identify which amino acids control the structural stability of that protein.
3. Amino acid participation factors. These provide a ranking of amino acids in terms of their footprint on the overall behaviour of a protein. Such information may be useful when it comes to designing new therapies.

The application of QCT on proteins generates data and information about structure, complexity, special arrangements of atoms etc. of the proteins. The knowledge, however, about the biological functions of the proteins derived from the above will have to be generated in collaboration with specialists from pharmaceutical R&D. This knowledge is crucial for designing new drugs.

References

1. Marczyk J (2009) A New Theory of Risk and Rating. Editrice Uniservice, Trento, Italy.
2. Giulio Molon, Jacek Marczyk, Giovanni Virzì, Riccardo Bellazzi, Alberto Malovini, et al. (2014) Analysis of ECG by means of Complexity Index and Association with Clinical Response to Cardiac Resynchronization Therapy". Journal of Cardiovascular Disease 2(2).
3. Giulio Molon, Francesco Solimene, Donato Melissano, Antonio Curnis, Giuseppina Belotti, et al. (2010) Baseline heart rate variability predicts clinical events in heart failure patients implanted with cardiac resynchronization therapy: validation by means of related complexity index. Ann Noninvasive Electro cardiol 15(4): 301-317.
4. Paweł Krzesiński, Jacek Marczyk, Bartosz Wolszczak, Grzegorz Gielerak (2021) Quantitative Complexity Theory Used in the Prediction of Head-Up Tilt Testing Outcome. Cardiol Res Pract: 8882498.
5. Paweł Krzesiński, Jacek Marczyk, Bartosz Wolszczak, Grzegorz Gerard Gielerak, Francesco Accardi (2023) Quantitative Complexity Theory (QCT) in Integrative Analysis of Cardiovascular Hemodynamic Response to Posture Change. Life 13(3): 632.
6. J Marczyk, P Stadler, JAF Treguerres (2023) Complexity Quantification and Comparison of Two Commercially Available Anti-Coagulants.Biomedical Journal of Scientific and Technical Research 49(2).
7. Giulio Molon, Jacek Marczyk, Giovanni Virzi, Francesco Accardi, A Costa, et al. (2013) ECG Predicts Response to Cardiac Resynchronization Therapy. Assessment by Means of Complexity Index. 44-th National Italian Cardiology Congress, Florence.
8. Batchinsky AI, et al. (2010) Changes in Systems-level Complexity Precede Deterioration in Traditional Vital Signs in Hypoxic Cardiac Arrest. American Heart Association Annual Meeting.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2023.50.007914

Marczyk J. Biomed J Sci & Tech Res