ÓMICAS

Julian Colorado Montaño
Investigador Principal Proyecto Ómicas
colorado@javeriana.edu.co

El futuro es de todos
Gobierno de Colombia
COLOMBIA CIENTÍFICA
Conocimiento Global para el Desarrollo

Geometric-Probabilistic Approach to Protein Secondary Structure Assingment

Sergio A. Quintana, Daniela Tamayo, Laura Gil, and , Carlos A. Arango
Department of Chemical Sciences
Universidad Icesi
Cali, Colombia

Simposio ÓMICAS 2022
Producción Sostenible y Seguridad Alimentaria desde las Ciencias ÓMICAS
Sesión 2: Proteómica y Metabolómica
Universidad Pontificia Javeriana
Cali, Colombia,
16 de Noviembre, 2022

LEVELS OF PROTEIN STRUCTURE

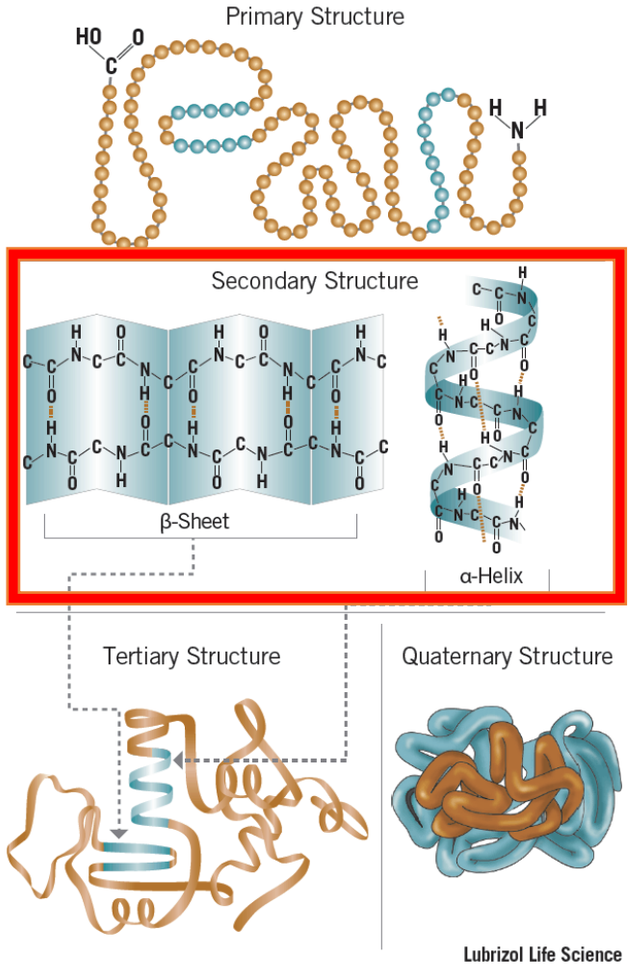


Figure 1. Protein structure levels.
Retreived from: https://lubrizolcdmo.com/technical-briefs/protein-structure/

Table 1. Transformation of secondary structure assignment from Q8 to Q3 the CASP reduction method.

Q3	Q8
Helix (H)	3/10-helix (G)
	α-helix (H)
Sheet (E)	β-sheet (E)
	β-bridge (B)
Coil (C)	π-helix (I)
	Turn (T)
	Bend (S)
	Loop (C)

METHODS TO DETERMINE PROTEIN'S STRUCTURE

Table 2. Examples of different methods for the determination of protein's structure

EXPERIMENTAL METHODS	COMPUTATIONAL METHODS	
• NMR • X-Ray Crystallography • cryo-EM	Prediction algorithms	Assignment algorithms

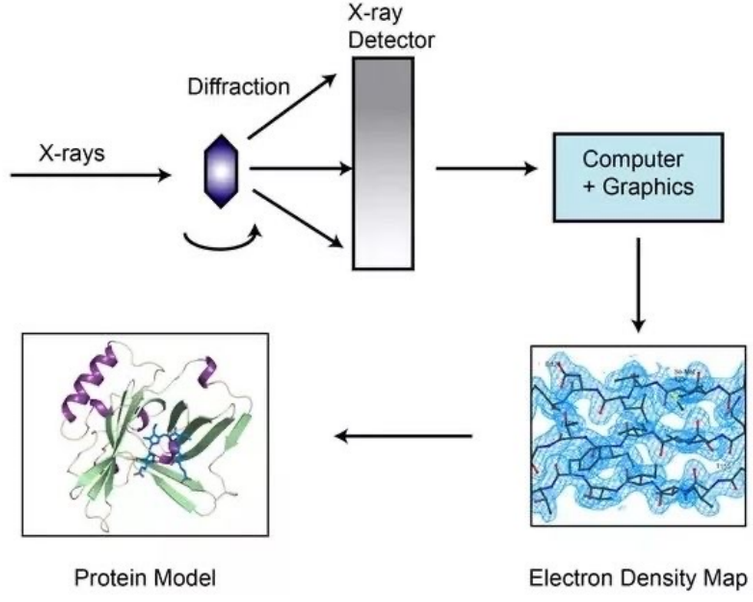


Figure 2. Methods for the determination of protein's structure.
Retreived from: https://www.creative-biolabs.com/protein-protein-interaction-assay-by-x-ray-crystallography-service.html

GEOMETRICAL ANALYSIS FOR AN ASSIGNMENT METHOD

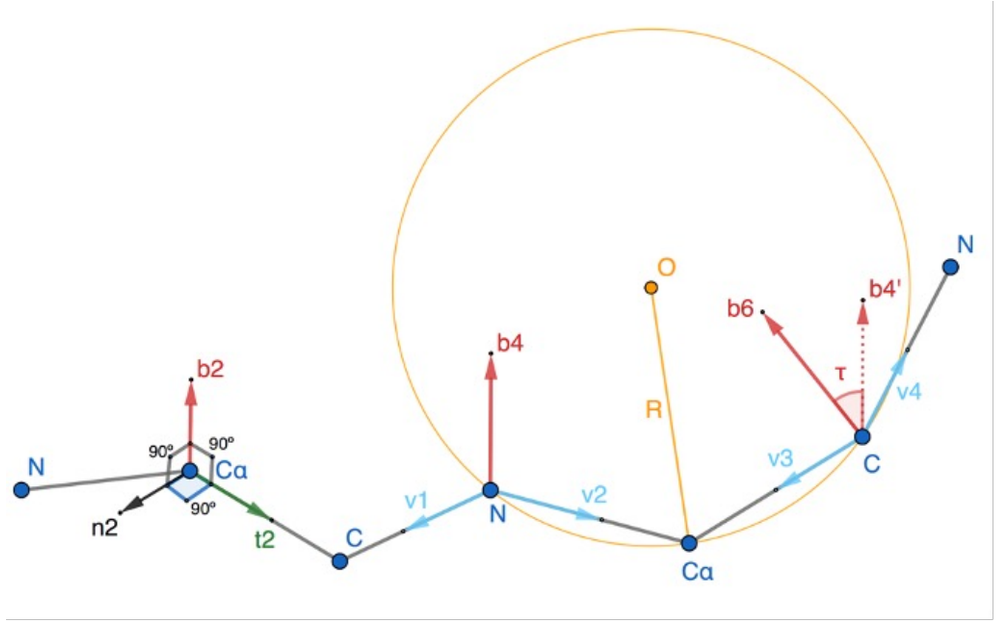


Figure 3. Geometrical definition of curvature and torsion in the backbone of a protein. Curvature (κ) defined as $\frac{1}{R}$ where R_i (in yellow) is the radius of the circumcircle between atoms N, C_α, and C in the i-th residue.

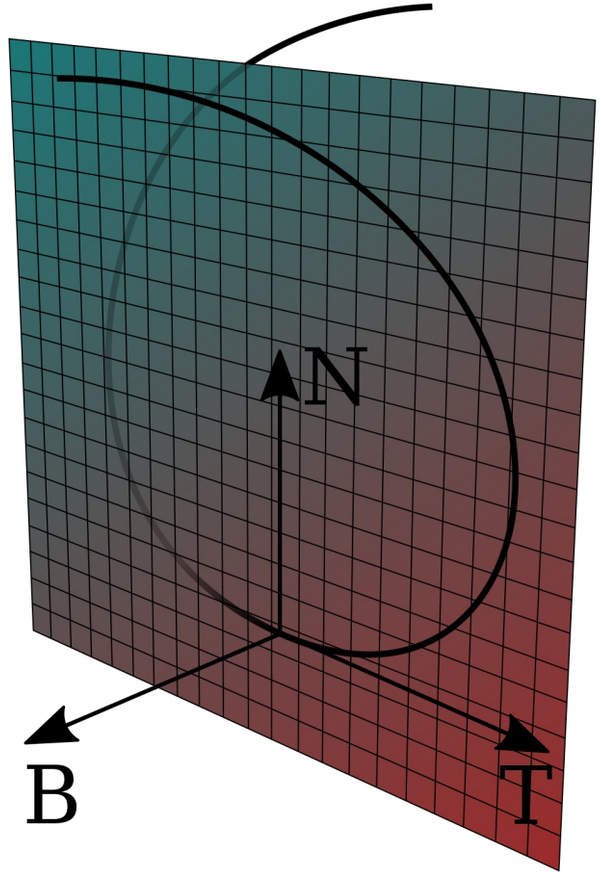


Figure 4. Frenet-Serret frame.
Retreived from: https://en.wikipedia.org/wiki/Frenet–Serret_formulas#/media/File:Frenet.svg

GEOMETRIC REPRESENTATION OF THE BACKBONE OF A PROTEIN

Proteins are large biomolecules with at least one chain of amino acid residues. Proteins can be structurally described at several levels. The primary structure of a single-chain protein P is the n-tuple formed by the n residue A_i of P,

$$\begin{aligned} P(\mathcal{P}) &= (\mathcal{A}_i)_{i=1}^{n} \\ &= (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n)\,. \end{aligned}$$

Each amino acid can be an element of the set of the 20 amino acids, $A_i \in G$, with

$$\begin{aligned} \mathbb{G} = \{&\mathrm{Ala}, \mathrm{Cys}, \mathrm{Asp}, \mathrm{Glu}, \mathrm{Phe}, \mathrm{Gly}, \mathrm{Ile} \\ &\mathrm{His}, \mathrm{Lys}, \mathrm{Leu}, \mathrm{Met}, \mathrm{Asn}, \mathrm{Pro}, \mathrm{Gln}, \\ &\mathrm{Arg}, \mathrm{Ser}, \mathrm{Thr}, \mathrm{Val}, \mathrm{Trp}, \mathrm{Tyr}\}. \end{aligned}$$

The amino acids have a chemical structure with a main chain of three bonded atoms (N,C_{α},C_{CO}) and a distinctive side chain R bonded to the C_{α} atom

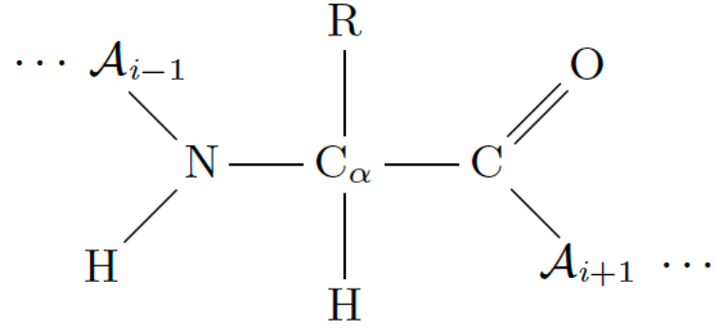


Figure 5. Chemical structural formula of the i-th residue, A_i, with side chain R, and main chain atoms (N,C_{α},C_{CO}).

GEOMETRIC REPRESENTATION OF THE BACKBONE OF A PROTEIN

The secondary structure of protein P is the n-tuple of secondary structure elements of the residues of P

$$S(\mathcal{P}) = (\sigma_i)_{i=1}^{n}$$

Secondary structure elements could be α-helices ($\sigma = \alpha$), β-sheets ($\sigma = \beta$), or coils ($\sigma = \gamma$). The value of σ_i is determined by structural methods. The backbone of P, $R(P)$, is the $3n$-tuple formed by the position vectors of the main chain atoms of the residues of P

$$R(\mathcal{P}) = (\boldsymbol{r}_i)_{i=1}^{3n}$$

The bonds of P, $T(P)$, form a $(3n-1)$-tuple with vectors $\boldsymbol{t}_i = \boldsymbol{r}_{i+1} - \boldsymbol{r}_i$,

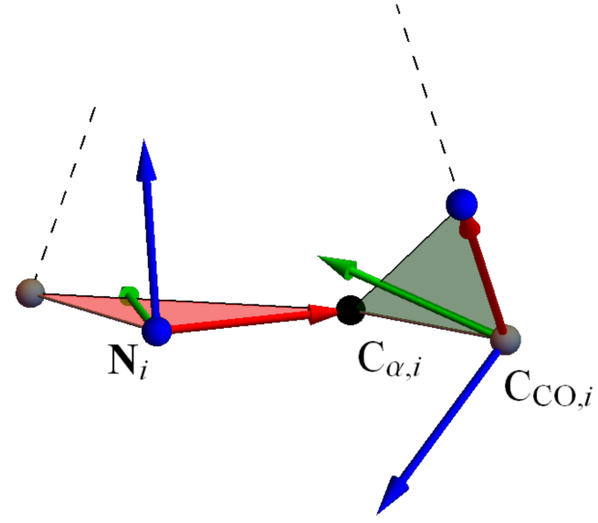
$$T(\mathcal{P}) = (\boldsymbol{t}_i)_{i=1}^{3n-1}$$

The binormals of P, $B(P)$, are the $(3n-2)$-tuple formed by the vectors $\mathbf{b}_i = \mathbf{t}_{i-1} \times \mathbf{t}_i$,

$$B(\mathcal{P}) = (\boldsymbol{b}_i)_{i=2}^{3n-1}$$

The normals of P, $N(P)$, is the $(3n-2)$-tuple formed by the vectors $\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i$,

$$N(\mathcal{P}) = (\boldsymbol{n}_i)_{i=2}^{3n-1}$$



CURVATURE AND TORSION OF THE BACKBONE OF A PROTEIN

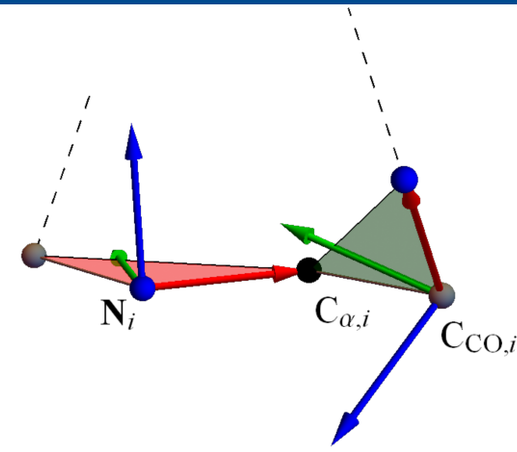


Figure 6. ($\mathbf{t}_i$,$\mathbf{n}_i$, $\mathbf{b}_i$) triad for the *i*-th residue of a protein. Bond vectors in red, normal vectors in green, and binormal vectors in blue.

The triad of position vectors $(r_j)_{j=i-1}^{i+1}$, with $i \in \{2, 3, \ldots, n-1\}$, circumscribes a circle C_i with center C_i and radius, ρ_i. The curvature of C_i, κ_i, is defined as $\kappa_i = \frac{1}{\rho_i}$. In this work, curvature is calculated only for the C_α atoms of the residues, which is given by

$$K_\alpha(\mathcal{P}) = (\kappa_{3i-1})_{i=1}^{n}$$

This work employs the dihedral angles, $\delta_{i,i+2}$, between the binormal vectors b_i and b_{i+2} of the N and the C_{CO} atoms, respectively, of the same residue. These angles are defined as torsions, and are given by the (n − 2)-tuple,

$$\tau(\mathcal{P}) = (\delta_{3i-2,3i})_{i=2}^{n-1}$$

CURVATURE AND TORSION OF THE BACKBONE OF A PROTEIN

These dihedrals are centered at the C_α atom of all the residues of P, except for the first and last residues. The pairing of the tuples $K_\alpha(P)$ and $\tau(P)$ gives the curvature-torsion pairs of protein P,

$$F(\mathcal{P}) = (f_i)_{i=2}^{n-1}$$

with $f_i = (\kappa_{3i-1}, \delta_{3i-2,3i})$. It is convenient to define the sets

$$\mathbf{F}_\alpha(\mathcal{P}) = \{f_i \in F(\mathcal{P}) : \sigma_i \in S(\mathcal{P}), \sigma_i = \alpha\},$$
$$\mathbf{F}_\beta(\mathcal{P}) = \{f_i \in F(\mathcal{P}) : \sigma_i \in S(\mathcal{P}), \sigma_i = \beta\},$$
$$\mathbf{F}_\gamma(\mathcal{P}) = \{f_i \in F(\mathcal{P}) : \sigma_i \in S(\mathcal{P}), \sigma_i = \gamma\}.$$

PROBABILITY DISTRIBUTIONS

Let P = $\{Pi\}_{i=1}^{m}$ be a set of *m* proteins. The sets

$$\mathbb{D}_{\alpha} = \bigcup_{i=1}^{m} \mathbf{F}_{\alpha}(\mathcal{P}_i),$$

$$\mathbb{D}_{\beta} = \bigcup_{i=1}^{m} \mathbf{F}_{\beta}(\mathcal{P}_i),$$

$$\mathbb{D}_{\gamma} = \bigcup_{i=1}^{m} \mathbf{F}_{\gamma}(\mathcal{P}_i),$$

PROTEIN DATABASE SELECTION OF 186 PROTEINS (SET OF PROTEINS).

All proteins chosen have known structure deposited un the RCSB Protein Data Bank (PDB).

CRITERIA

- ✓ The structure obtaining method must has coordinates of its atoms.
- ✓ The chain length should be between 50-700 aa.
- ✓ Proteins belonging to any organism.
- ✓ Any type of protein classification is allowed
- ✓ If the protein has more than one chain, only the first one will be taken.
- ✓ Proteins with mutations were not taken.

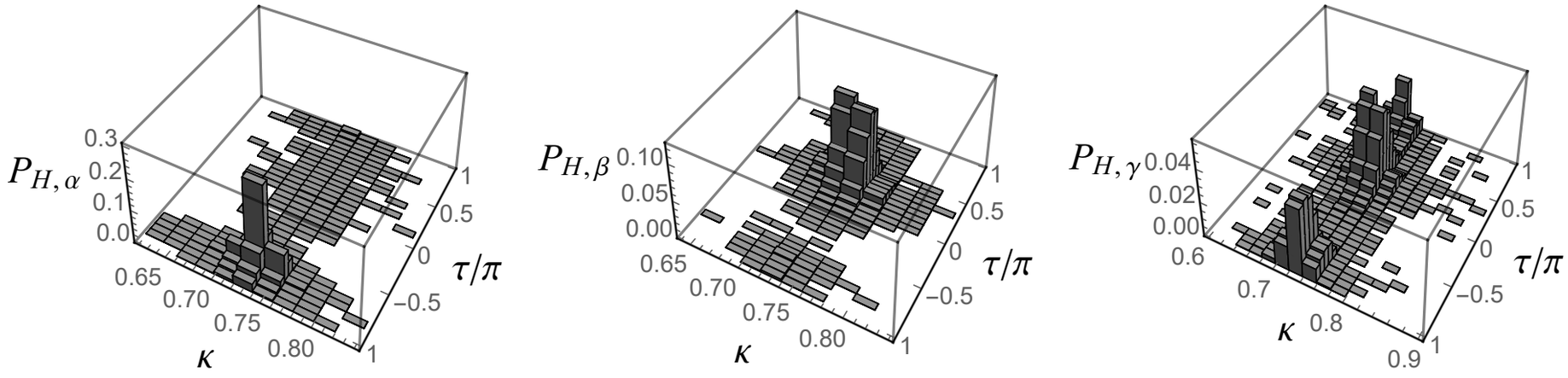


Figure 7. Probability histograms for $P_{H,\alpha}$, $P_{H,\beta}$, and $P_{H,\gamma}$ as functions of κ and τ .

Region's delimitations

$$\mathcal{R}_\alpha = \{(\kappa, \tau) : P_\alpha(\kappa, \tau) > 0.50\},$$
$$\mathcal{R}_\beta = \{(\kappa, \tau) : P_\beta(\kappa, \tau) > 0.45\},$$
$$\mathcal{R}_\gamma = \{(\kappa, \tau) : P_\gamma(\kappa, \tau) > 0.50\},$$

DISTRIBUTION OF CURVATURE AND TORSION VALUES KTS²

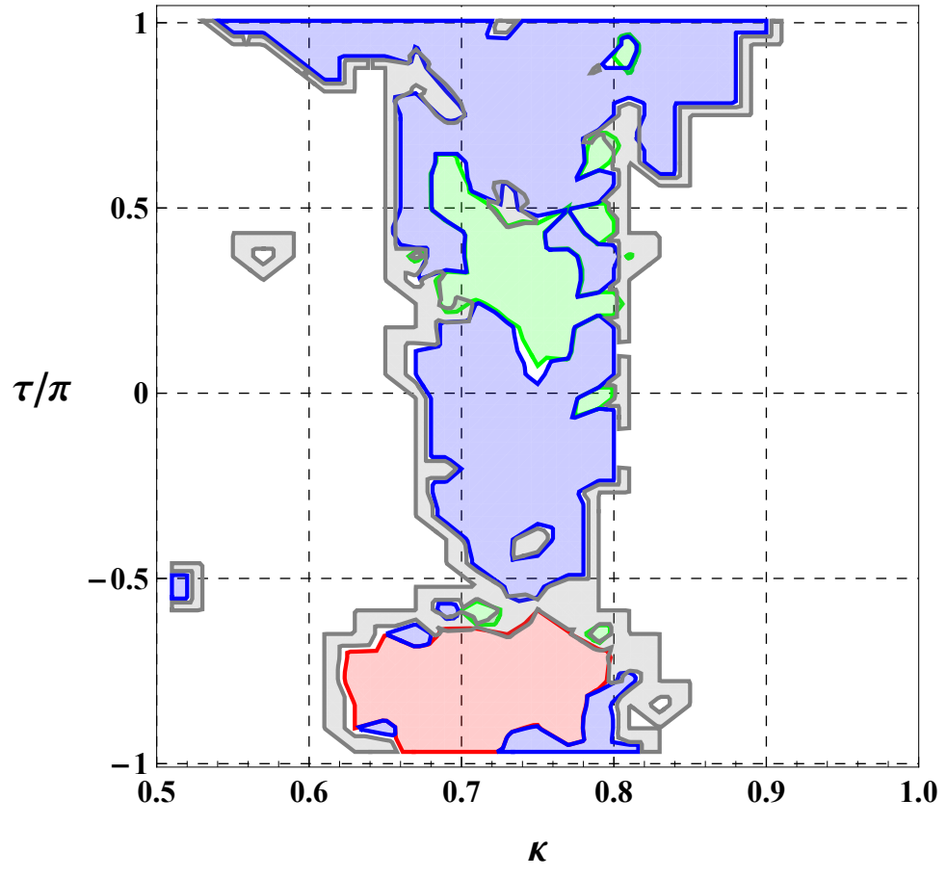


Figure 8. Secondary structure probability regions on the κτ-plane. α-helix region R_α in red, β-sheet región R_β in green, and γ-coil region R_γ in blue.

$$\Sigma(\kappa,\tau) = \begin{cases} \alpha, & (\kappa,\tau) \in \mathcal{R}_\alpha, \\ \beta, & (\kappa,\tau) \in \mathcal{R}_\beta, \\ \gamma, & (\kappa,\tau) \in \mathcal{R}_\gamma. \end{cases}$$

EVALUATION OF THE PERFORMANCE OF KTS^2 METHOD

Table 2. Percentage of correctly assigned secondary structure (SS) by different assignment tools for the set P using the SS reported by the author of the PDB file as reference.

	Author SS in comparison with			
	DSSP	**STRIDE**	**SST**	**KTS^2**
QH	82.58	85.02	74.18	80.08
QE	94.66	94.04	77.46	76.04
QC	94.14	90.26	75.12	42.15
Q3	91.13	90.96	78.73	66.95

Table 3. Performance of different secondary structure assignment tools for 29 testing proteins, compared to the secondary structure reported by the author of the PDB file.

	Author SS in comparison with			
	DSSP	**STRIDE**	**SST**	**KTS^2**
QH	72.74	74.31	67.06	78.18
QE	100.0	97.83	82.01	78.26
QC	95.79	90.35	72.13	38.67
Q3	91.81	90.68	79.67	63.79

APPLICATION OF KTS^2 METHOD TO GCR1 PROTEIN

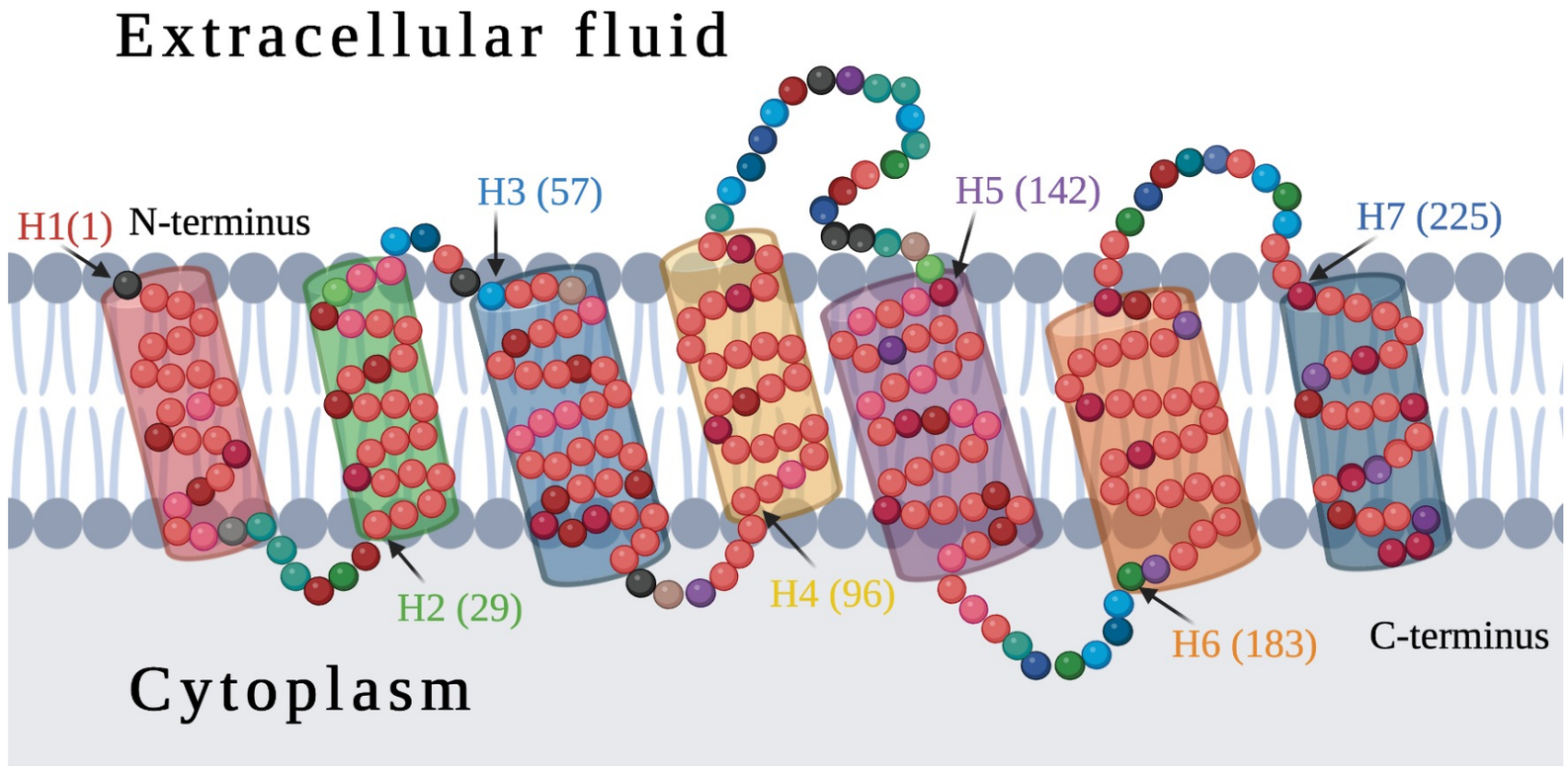


Figure 9. Extended representation of the secondary structure of GCR1. This transmembrane protein has seven helices. Each dot represents a residue of GCR1. The color of the dots is assigned by using the probabilies of κτS^2 and the RGB color scale.

APPLICATION OF KTS^2 METHOD TO GCR1 PROTEIN

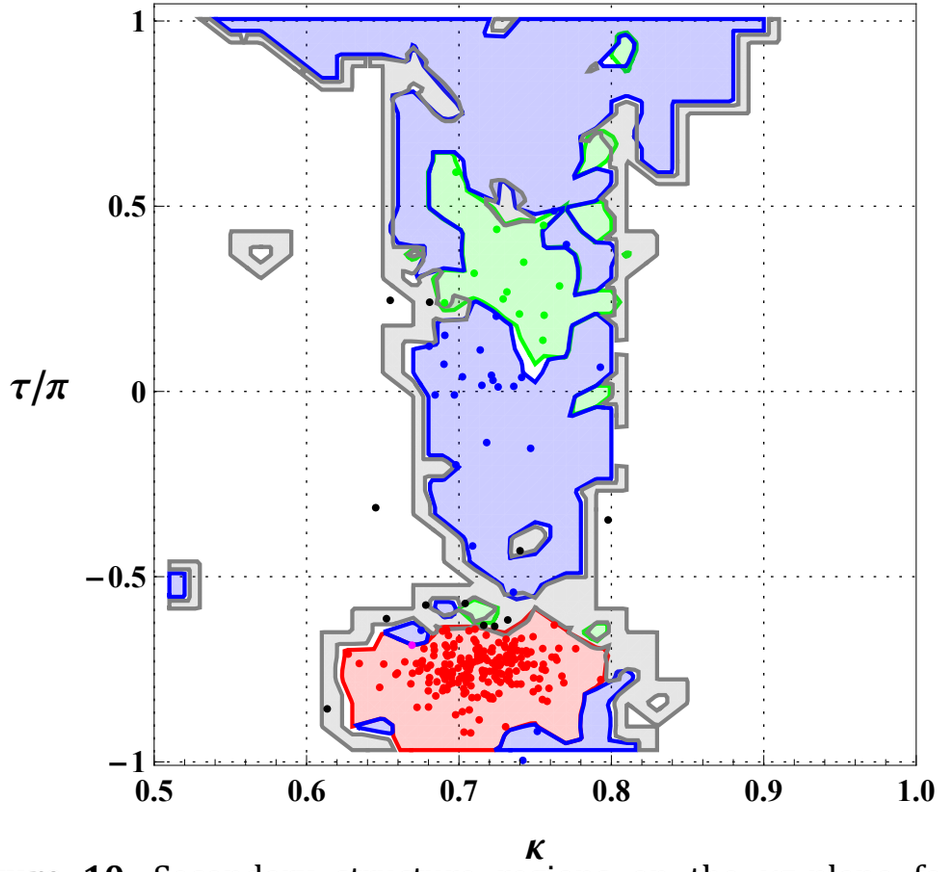


Figure 10. Secondary structure regions on the κτ-plane for GCR1 protein. Colored dots represent the assignment of a secondary structure element for a given residue: α-helix in red, β-sheet in green, and γ-coil in blue. Black dots are assumed as residues with a not clearly differentiable representation of their secondary structure by κτS^2.

Table 4. Successful assignment performance of κτS^2 in GCR1 in comparison with STRIDE.

	STRIDE	KTS^2
QH	96.86	94.24
QE	Indeterminate	Indeterminate
QC	77.59	37.93
Q3	92.37	81.13

CONCLUSIONS AND PERSPECTIVES

- The $\kappa\tau S^2$ method works as an alternative way to distinguish helices, sheets, and coils as structures with defined curvature and torsion patterns, instead of seeing them as hydrogen bond-patterned-defined structures.
- The existing error by $\kappa\tau S^2$ in the assignment of helices, sheets, coils, and the overall agreement error, is related to the overlapping between regions in the secondary structure $\kappa\tau$ plane.
- The small size of set P may not be large enough to achieve the statistical robustness necessary to allow a better separation of some of the regions of the probability $\kappa\tau$ plane and higher values of QH, QE, and QC.
- The use of additional geometrical descriptors would give rise to new density and contour maps that would improve the probability functions P_α, P_β, and P_γ.

REFERENCES

D. Eisenberg, Proceedings of the National Academy of Sciences 100, 11207 (2003).
C. BrÅNandÅLen and J. Tooze, in Bioinformatics for geneticists, edited by C. BrÅNandÅLen and Tooze (Garland Pub, New York, 1999) pp. 14–19.
J. Ludwiczak, A. Winski, A. da Silva Neto, and et al., Sci Rep 9, 6888 (2019).
T. Smolarczyk, T. Roterman-Konieczna, and K. Stapor, Current Bioinformatics 15, 90 (2020).
C. Lee, Endocrinology and Metabolism 32, 18 (2017).
Y. Xu, D. Xu, and J. Liang, Computational Methods for Protein Structure Prediction and Modeling: Volume 2: Structure Prediction (Springer, New York, 2010).
S. Hu, “Dynamics of discrete curves with applications to protein structure,” (2013), acta Universitatis Upsaliensis https://uu.diva-portal.org/smash/get/diva2:621923/FULLTEXT01.pdf.

ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS

Semillero de Fisicoquímica Teórica – Grupo $\kappa\tau S^2$

Daniela Tamayo

Laura Gil

Sergio A. Quintana

Carlos A. Arango


ÓMICAS

ÓMICAS

Aliados

CIAT
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

cenicaña

FEDEARROZ
FEDERACIÓN NACIONAL DE ARROCEROS

NIAB

intelecto
Soluciones y Tecnología

HITECH
AUTOMATIZACION S.A.

Caltech

GHENT UNIVERSITY

東京大学
THE UNIVERSITY OF TOKYO

ILLINOIS

Pontificia Universidad JAVERIANA Cali
IES Ancla

UNIVERSIDAD DEL QUINDÍO

Universidad de Ibagué

UNILLANOS
UNIVERSIDAD DE LOS LLANOS

Universidad de los Andes
Colombia

UNIVERSIDAD ICESI

Pontificia Universidad JAVERIANA Bogotá

Apoyan

El futuro es de todos
Gobierno de Colombia

COLOMBIA CIENTÍFICA
Conocimiento Global para el Desarrollo

ÓMICAS

Julian Colorado Montaño
Investigador Principal Proyecto Ómicas
colorado@javeriana.edu.co

El futuro es de todos
Gobierno de Colombia
COLOMBIA CIENTÍFICA
Conocimiento Global para el Desarrollo