Motif alignment and protein secondary structure

Čutura, Ivan

UNIVERSITY OF ZAGREB

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS

Ivan Čutura

MOTIF ALIGNMENT AND PROTEIN SECONDARY STRUCTURE

Diploma thesis

Supervisor of the thesis:
doc. dr. sc. Pavle Goldstein

Zagreb, July, 2024

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Želim izraziti svoju duboku zahvalnost mentoru, doc. dr. sc. Pavlu Goldsteinu, na strpljenju i pomoći pri pisanju ovog rada. Posebno zahvaljujem svojim roditeljima na neograničenoj podršci. Također, hvala mojim prijateljima i bratu, a osobito tebi, Antonela, na tvojoj neizmjernoj motivaciji i ohrabrenju.

Table of Contents

Introduction

Motif scanning is a common method in bioinformatics, used for various purposes such as protein family assignment, secondary structure prediction, and similar tasks. Motif scanning methods take a motif, which we call a query, as input and search for similar patterns in a set of sequences. The output consists of sufficiently similar matches, forming a set of positives, or what is referred to as the response. So, the objective of the motif scanning method is to detect motifs of sufficient similarity to the query, which is then used to determine family membership or structural and functional features or assignments.

In this thesis, we are focused on improving the accuracy of motif scanning procedures. Given a set of motifs obtained from a scanning process, we filter the response based on the analysis of pairwise similarity of the motifs. We achieve this by transitioning to Euclidean space, where we consider the distance between motifs instead of their similarity.

This work consists of five chapters. The first chapter covers mathematical concepts from linear algebra, probability, and statistics that are important later on. The second chapter explains biological concepts and terms from bioinformatics. The third and fourth chapters provide a detailed description of the method for improving the accuracy of the motif scanning procedure when dealing with secondary structure motifs. Finally, in the last chapter, we demonstrate the method in detail with an alpha helix example. At the end we present the results that we obtained.

Chapter 1

Mathematical Background

Theorems, definitions, propositions, and remarks on linear algebra and probability in this chapter are taken from sources [4], [5], [6], [12] and [14].

1.1 Linear algebra

Definition 1.1.1. *Let $\mathbb{F}$ be a set with the binary operations of addition $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ and multiplication $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ that satisfy the following properties:*

1. $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma, \quad \forall \alpha, \beta, \gamma \in \mathbb{F}$;
2. *There exists* $0 \in \mathbb{F}$ *such that* $\alpha + 0 = 0 + \alpha = \alpha, \quad \forall \alpha \in \mathbb{F}$;
3. *For every* $\alpha \in \mathbb{F}$, *there exists* $-\alpha \in \mathbb{F}$ *such that* $\alpha + (-\alpha) = -\alpha + \alpha = 0$;
4. $\alpha + \beta = \beta + \alpha, \quad \forall \alpha, \beta \in \mathbb{F}$;
5. $\alpha(\beta\gamma) = (\alpha\beta)\gamma, \quad \forall \alpha, \beta, \gamma \in \mathbb{F}$;
6. *There exists* $1 \in \mathbb{F} \setminus \{0\}$ *such that* $1 \cdot \alpha = \alpha \cdot 1 = \alpha, \quad \forall \alpha \in \mathbb{F}$;
7. *For every* $\alpha \in \mathbb{F}$, $\alpha \neq 0$, *there exists* $\alpha^{-1} \in \mathbb{F}$ *such that* $\alpha\alpha^{-1} = \alpha^{-1}\alpha = 1$;
8. $\alpha\beta = \beta\alpha, \quad \forall \alpha, \beta \in \mathbb{F}$;
9. $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, \quad \forall \alpha, \beta, \gamma \in \mathbb{F}$.

*Then we say that the ordered triple $(\mathbb{F}, +, \cdot)$ is a **field**, and the elements of the field are called **scalars**.*

Remark 1.1.2. *The set of real numbers* $\mathbb{R}$ *with the usual operations of addition and multiplication is a field.*

Definition 1.1.3. *Let* V *be a non-empty set with a binary operation of addition* $+ : V \times V \to V$ *and a scalar multiplication operation where scalars are from the field* $\mathbb{F}$, $\cdot : \mathbb{F} \times V \to V$. *We say that the ordered triple* $(V, +, \cdot)$ *is a* ***vector space*** *over the field* $\mathbb{F}$ *if the following hold:*

1. $a + (b + c) = (a + b) + c, \quad \forall a, b, c \in V;$
2. *There exists* $0 \in V$ *such that* $a + 0 = 0 + a = a, \quad \forall a \in V;$
3. *For every* $a \in V$, *there exists* $-a \in V$ *such that* $a + (-a) = -a + a = 0;$
4. $a + b = b + a, \quad \forall a, b \in V;$
5. $\alpha(\beta a) = (\alpha\beta)a, \quad \forall \alpha, \beta \in \mathbb{F}, \forall a \in V;$
6. $(\alpha + \beta)a = \alpha a + \beta a, \quad \forall \alpha, \beta \in \mathbb{F}, \forall a \in V;$
7. $\alpha(a + b) = \alpha a + \alpha b, \quad \forall \alpha \in \mathbb{F}, \forall a, b \in V;$
8. $1 \cdot a = a, \quad \forall a \in V.$

Remark 1.1.4. *The set* $\mathbb{R}^n$ *with the usual operations of addition and scalar multiplication is a vector space over the field* $\mathbb{R}$. *We also say that* $(\mathbb{R}^n, +, \cdot)$ *is a* ***real*** *vector space.*

Definition 1.1.5. *For natural numbers* m *and* n, *mapping*

$$A : \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\} \to \mathbb{F}$$

is called a ***matrix*** *of type* (m, n) *with coefficients from the field* $\mathbb{F}$.

Definition 1.1.6. *Let* V *be a vector space over the field* $\mathbb{F}$. *A* ***scalar product*** *on* V *is a mapping* $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ *that satisfies the following properties:*

1. $\langle x, x \rangle \geq 0, \quad \forall x \in V;$
2. $\langle x, x \rangle = 0 \Leftrightarrow x = 0;$
3. $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle, \quad \forall x_1, x_2, y \in V;$
4. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle, \quad \forall \alpha \in \mathbb{F}, \forall x, y \in V;$
5. $\langle x, y \rangle = \langle y, x \rangle, \quad \forall x, y \in V.$

Remark 1.1.7. *In $\mathbb{R}^n$, the canonical scalar product is defined by*

$$\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i y_i.$$

Definition 1.1.8. *A vector space on which a scalar product is defined is called a* ***unitary space****.*

Definition 1.1.9. *Let V be a unitary space. A* ***norm*** *on V is a function $\|\cdot\| : V \to \mathbb{R}$ defined by*

$$\|x\| = \sqrt{\langle x, x \rangle}.$$

Proposition 1.1.10. *The norm on a unitary space V has the following properties:*

1. $\|x\| \geq 0, \quad \forall x \in V$;
2. $\|x\| = 0 \Leftrightarrow x = 0$;
3. $\|\alpha x\| = |\alpha| \|x\|, \quad \forall \alpha \in \mathbb{F}, \forall x \in V$;
4. $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in V$.

Remark 1.1.11. *Any function $\|\cdot\| : V \to \mathbb{R}$ on a vector space V with the properties from Proposition 1.1.10 is called a norm. Then $(V, \|\cdot\|)$ is called a* ***normed space****.*

Remark 1.1.12. *The norm induced by the canonical scalar product on $\mathbb{R}^n$, defined in Remark 1.1.7, is given by the formula*

$$\|(x_1, \ldots, x_n)\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

This norm is called the ***Euclidean norm****.*

Definition 1.1.13. *Let V be a normed space. A* ***metric*** *or* ***distance*** *on the set V is a mapping $d : V \times V \to \mathbb{R}$ defined by*

$$d(x, y) = \|x - y\|.$$

Proposition 1.1.14. *The metric on a normed space has the following properties:*

1. $d(x, y) \geq 0, \quad \forall x, y \in V$;
2. $d(x, y) = 0 \Leftrightarrow x = y, \quad \forall x, y \in V$;
3. $d(x, y) = d(y, x), \quad \forall x, y \in V$;

4. $d(x, y) \leq d(x, z) + d(z, y), \quad \forall x, y, z \in V.$

Remark 1.1.15. *Let X be a non-empty set. Any function $d : X \times X \to \mathbb{R}$ on the set X with the properties from Proposition 1.1.14 is called a metric or distance. Then (X, d) is called a* ***metric space****.*

Remark 1.1.16. *The metric induced by the Euclidean norm on $\mathbb{R}^n$, defined in Remark 1.1.12, is given by the formula*

$$d\left((x_1, \ldots, x_n), (y_1, \ldots, y_n)\right) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

This metric is called the ***Euclidean metric****, and the space $\mathbb{R}^n$ together with this metric is called an* ***n-dimensional Euclidean space****.*

Definition 1.1.17. *Let (X, d) be a metric space, and let $a \in X$ and $r \in \mathbb{R}$, $r > 0$. The set*

$$K(a, r) = \{x \in X \mid d(a, x) < r\},$$

is called an ***open ball*** *in X with center at a and radius r.*

Remark 1.1.18. *In an n-dimensional Euclidean space $\mathbb{R}^n$, an open ball with center at a and radius r is given by*

$$K(a, r) = \left\{ x \in \mathbb{R}^n \mid \sqrt{\sum_{i=1}^{n} (a_i - x_i)^2} < r \right\}.$$

1.2 Probability theory

Probability Space

Definition 1.2.1. *A* ***random experiment****, or a* ***random trial****, is an experiment whose outcomes, i.e., results, are not uniquely determined by the conditions under which we conduct the experiment.*

Definition 1.2.2. *The* ***sample space*** Ω *is a non-empty set that represents the set of all outcomes of a random experiment. The elements ω of the set Ω are called* ***elementary events****.*

Definition 1.2.3. *A family $\mathcal{F}$ of subsets of Ω ($\mathcal{F} \subset \mathcal{P}(\Omega)$) is a* ***$\sigma$-algebra of sets*** *on Ω if:*

1. $\emptyset \in \mathcal{F}$;

2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;

3. $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definition 1.2.4. *Let $\mathcal{F}$ be a σ-algebra on the set Ω. The ordered pair $(\Omega, \mathcal{F})$ is called a* ***measurable space****.*

Definition 1.2.5. *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is a* ***probability*** *(on $\mathcal{F}$, on Ω) if it satisfies:*

1. $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$;

2. $\mathbb{P}(\Omega) = 1$;

3. $A_i \in \mathcal{F}, i \in \mathbb{N}$ *and* $A_i \cap A_j = \emptyset$ *for* $i \neq j \Longrightarrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition 1.2.6. *An ordered triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra on Ω and $\mathbb{P}$ is a probability on $\mathcal{F}$, is called a* ***probability space****.*

Random variable

Definition 1.2.7. *Let S be an arbitrary non-empty set and $\mathcal{A}$ be a family of subsets of S ($\mathcal{A} \subset \mathcal{P}(S)$). Denote by $\sigma(\mathcal{A})$ the smallest σ-algebra of subsets of S containing $\mathcal{A}$. We call it the* ***σ-algebra generated by*** *$\mathcal{A}$.*

Definition 1.2.8. *Let $\mathcal{B}$ denote the σ-algebra generated by the family of all open sets on $\mathbb{R}$. $\mathcal{B}$ is called the* ***Borel σ-algebra*** *on $\mathbb{R}$, and the elements of the σ-algebra $\mathcal{B}$ are called* ***Borel sets*** *.*

Definition 1.2.9. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is a* ***random variable*** *(on Ω) if $X^{-1}(B) \in \mathcal{F}$ for arbitrary $B \in \mathcal{B}$, i.e., $X^{-1}(B) \subset \mathcal{F}$.*

Definition 1.2.10. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}^n$. We say that X is an* ***n-dimensional random vector*** *(or simply a* ***random vector****) (on Ω) if $X^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{B}^n$, i.e., $X^{-1}(\mathcal{B}^n) \subset \mathcal{F}$.*

Definition 1.2.11. *Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. X is a* ***simple random variable*** *if its range is a finite set.*

X is a simple random variable if and only if

$$X = \sum_{k=1}^{n} x_k \mathcal{K}_{A_k}$$

where $x_1, x_2, \ldots, x_n$ are real numbers, and $A_1, A_2, \ldots, A_n$ are pairwise disjoint events with $\bigcup_{k=1}^{n} A_k = \Omega$. $\mathcal{K}_{A_k}$ denotes the characteristic function of the set A_k.

Let $X_1, X_2 : \Omega \to \mathbb{R}$. Then we define the functions $X_1 \vee X_2$ and $X_1 \wedge X_2$ on Ω by:

$$(X_1 \vee X_2)(\omega) = \max\{X_1(\omega), X_2(\omega)\}, \quad \omega \in \Omega,$$

and

$$(X_1 \wedge X_2)(\omega) = \min\{X_1(\omega), X_2(\omega)\}, \quad \omega \in \Omega.$$

Using first of the two functions, we define the positive and negative parts of the real function X on Ω:

$$X^+ = X \vee 0, X^- = (-X) \vee 0.$$

X^+ and X^- are non-negative real functions, and we have:

$$X = X^+ - X^-,$$

$$|X| = X^+ + X^-.$$

Corollary 1.2.12. *X is a random variable if and only if X^+ and X^- are random variables.*

Theorem 1.2.13. *Let X be a non-negative random variable on Ω. Then there exists an increasing sequence $(X_n, n \in \mathbb{N})$ of non-negative simple random variables such that $X = \lim_{n\to\infty} X_n$ (on Ω).*

Mathematical expectation and variance

The definition of mathematical expectation is conducted in three steps. First, the mathematical expectation of a simple random variable is defined, then of a non-negative random variable, and finally of a general random variable.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{K}$ be the set of all simple random variables defined on Ω, and $\mathcal{K}_+$ the set of all non-negative functions in $\mathcal{K}$.

Let $X \in \mathcal{K}$, $X = \sum_{k=1}^{n} x_k \mathcal{K}_{A_k}$, where $A_1, A_2, \ldots, A_n \in \mathcal{F}$ are mutually disjoint.

Definition 1.2.14. ***Mathematical expectation*** *of X, or simply the* ***expectation*** *of X, is denoted by $\mathbb{E}[X]$ and defined as:*

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathbb{P}(A_k).$$

Now let X be a **non-negative random variable** defined on Ω. According to Theorem 1.2.13, there exists an increasing sequence $(X_n)_{n\in\mathbb{N}}$ of non-negative simple random variables such that $X = \lim_{n\to\infty} X_n$. The sequence $(\mathbb{E}[X_n])_{n\in\mathbb{N}}$ is an increasing sequence in $\mathbb{R}_+$, so $\lim_{n\to\infty} E[X_n]$ exists and may be equal to $+\infty$.

Definition 1.2.15. ***Mathematical expectation*** *of X, or simply the* ***expectation*** *of X, is defined as*

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

Now let X be an **arbitrary random variable** on Ω. It holds that $X = X^+ - X^-$, where X^+ and X^- are non-negative random variables and $X^+, X^- \geq 0$.

Definition 1.2.16. *We say that the* ***mathematical expectation*** *of X, or simply the* ***expectation*** *of X,* ***exists*** *(or is defined) if at least one of the quantities $\mathbb{E}[X^+]$, $\mathbb{E}[X^-]$ is finite, i.e., if $\min\{\mathbb{E}[X^+], \mathbb{E}[X^-]\} < +\infty$. Then by definition, we set*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

We list basic properties of mathematical expectation:

Theorem 1.2.17. *We have:*

1. If $\mathbb{E}[X]$ exists and $c \in \mathbb{R}$, then $\mathbb{E}[cX]$ exists and
$$\mathbb{E}[cX] = c\mathbb{E}[X].$$
2. If $X \leq Y$, then
$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$
In the sense that
$$\text{if } -\infty < \mathbb{E}[X], \text{ then } -\infty < \mathbb{E}[Y] \text{ and } \mathbb{E}[X] \leq \mathbb{E}[Y],$$
or
$$\text{if } \mathbb{E}[Y] < \infty, \text{ then } \mathbb{E}[X] < \infty \text{ and } \mathbb{E}[X] \leq \mathbb{E}[Y].$$
3. If $\mathbb{E}[X]$ exists, then
$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$
4. If $\mathbb{E}[X]$ exists, then $\mathbb{E}[X\mathcal{K}_A]$ exists for every $A \in \mathcal{F}$. If $\mathbb{E}[X]$ is finite, then $\mathbb{E}[X\mathcal{K}_A]$ is finite for every $A \in \mathcal{F}$.
5. Let X and Y be non-negative random variables. Then
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Definition 1.2.18. *Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathbb{E}[X]$ be finite. Then we define the* ***variance*** *of X, denoted by $\mathrm{Var}(X)$ or σ_X^2, as follows:*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Remark 1.2.19. *The positive square root of the variance is called the* ***standard deviation*** *of X and is denoted by σ_X.*

Distribution function

Definition 1.2.20. *Let X be a random variable on Ω. The **distribution function of** X is the function $F_X : \mathbb{R} \to [0,1]$, defined as:*

$$F_X(x) = \mathbb{P}(X^{-1}(\langle -\infty, x])) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

Remark 1.2.21. *If it is clear which random variable we are referring to, we write F instead of F_X.*

Theorem 1.2.22. *The distribution function F of the random variable X is non-decreasing and right-continuous on $\mathbb{R}$, and satisfies:*

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

*A function $F : \mathbb{R} \to [0,1]$ with these properties is called the **cumulative distribution function** (on $\mathbb{R}$) or simply, the **distribution function**.*

Classification of random variables

Definition 1.2.23 (Discrete Random Variable)**.** *Let X be a random variable on Ω. X is **discrete** if there exists a finite or countable set $D \subset \mathbb{R}$ such that $\mathbb{P}\{X \in D\} = 1$.*

Discrete random variables are typically specified by providing the set $D = \{x_1, x_2, \ldots\}$ and the probabilities $p_n = \mathbb{P}\{X = x_n\}$, which can be represented in tabular form:

$$X \sim \begin{pmatrix} x_1 & x_2 \ldots & x_n \ldots \\ p_1 & p_2 \ldots & p_n \ldots \end{pmatrix}.$$

The above table is called the **distribution** of the random variable X. In the distribution table, $x_n \in \mathbb{R}$, $x_i \neq x_j$ for $i \neq j$, $p_n > 0$, and $\sum_n p_n = 1$.

Definition 1.2.24. *A function $g : \mathbb{R} \to \mathbb{R}$ is a **Borel function** if $g^{-1}(B) \in \mathcal{B}$ for every $B \in \mathcal{B}$, i.e., if $g^{-1}(\mathcal{B}) \subset \mathcal{B}$.*

Definition 1.2.25. *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let F_X denote its cumulative distribution function. X is **absolutely continuous**, or simply, **continuous random variable**, if there exists a non-negative real-valued Borel function f on $\mathbb{R}$($f : \mathbb{R} \to \mathbb{R}_+$) such that*

$$F_X(x) = \int_{-\infty}^{x} f(t)\, d\lambda(t), \quad x \in \mathbb{R}.$$

If X is a continuous random variable, the function f is called the ***probability density function*** *of X, denoted sometimes as f_X.*

Definition 1.2.26. *Let $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$. A continuous random variable X has a* ***normal distribution with parameters*** *μ and σ^2 if its density function f is given by*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

We denote this as $X \sim N(\mu, \sigma^2)$.

Remark 1.2.27. *X is the* ***standard normal distribution*** *if $X \sim N(0, 1)$, hence the probability density function $f(x)$ is given by*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

Descriptive statistics

In this section, we define the terms that are necessary for further understanding of the paper. We introduce concepts such as the arithmetic mean, sample standard deviation, sample variance, and data standardization.

Let $x_1, x_2, \ldots, x_n$ be n values (observations) of the variable X, comprising a dataset. If X is a numerical variable, it is a sequence of numbers. We assume that X is a numerical variable.

The **arithmetic mean** of the data is a measure of central tendency defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The **sample variance** is a measure of data dispersion, representing the average squared deviation of the data points from their arithmetic mean, given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

From the previous definitions, the **sample standard deviation**, which is the square root of the variance, is given by:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Data standardization is a common procedure in statistics before processing data and building models or algorithms. Data are transformed by subtracting the mean and dividing by the sample standard deviation:

$$x_i' = \frac{x_i - \bar{x}}{s}.$$

Chapter 2

Biological Background and Bioinformatics

Proteins and their structures

Proteins

Proteins are essential biomolecules composed of amino acids. They serve as the building blocks of cells, tissues, and organs, participating in virtually every physiological function in living organisms. Amino acids are organic molecules that contain both an amino group (NH2) and a carboxyl group (COOH). Each amino acid has a specific side chain, known as an R group, that defines its characteristics.

Label	**Name**	**Label**	**Name**
A	Alanine	M	Methionine
C	Cysteine	N	Asparagine
D	Aspartic Acid	P	Proline
E	Glutamic Acid	Q	Glutamine
F	Phenylalanine	R	Arginine
G	Glycine	S	Serine
H	Histidine	T	Threonine
I	Isoleucine	V	Valine
K	Lysine	W	Tryptophan
L	Leucine	Y	Tyrosine

Table 2.1: Standard amino acids: labels and names

Proteins are formed by linking amino acids together in a linear chain through peptide

bonds. This chain has two ends: the N-terminus (amino terminus), which has a free amino group, and the C-terminus (carboxyl terminus), which has a free carboxyl group. The sequence of amino acids in a protein, read from the N-terminus to the C-terminus, determines the protein's unique structure and function [15].

Proteins exhibit diverse functions. They act as enzymes, structural components, signaling molecules, gene expression regulators, and more. The specific function of a protein is linked to its three-dimensional conformation, which is determined by the amino acid sequence. Understanding protein structure and function is crucial for cellular insights and therapeutic development.

Protein structure

Protein structures are categorized into primary, secondary, tertiary, and quaternary. The primary structure is the linear sequence of amino acids connected by peptide bonds, which dictates the folding and formation of the protein's secondary structure. The secondary structure consists of local folding patterns of the backbone that are stabilized by hydrogen bonds. The tertiary structure refers to the overall three-dimensional shape of a single polypeptide chain. The quaternary structure is the assembly of multiple polypeptide chains into a functional complex.

Secondary structure

As mentioned, the secondary structure consists of local folding patterns of the backbone that are stabilized by hydrogen bonds between nearby amino acids in the polypeptide chain. There are two main types of secondary structures: α-helices and β-sheets. The α-helix is a right-handed helical structure formed by the twisting of the polypeptide chain into a coil-like shape, and it is the most common structural arrangement in the secondary structure. The β-sheet consists of strands of polypeptide chains that are extended and aligned alongside each other, with adjacent strands held together by hydrogen bonds. β-sheets can be antiparallel or parallel, depending on the relative direction of the polypeptide chains [16].

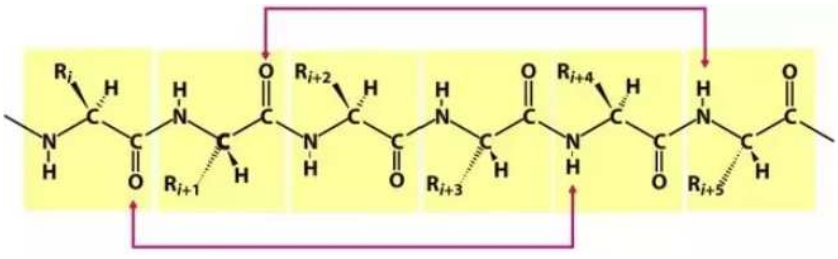

Figure 2.1: Formation of hydrogen bonds in the α-helix

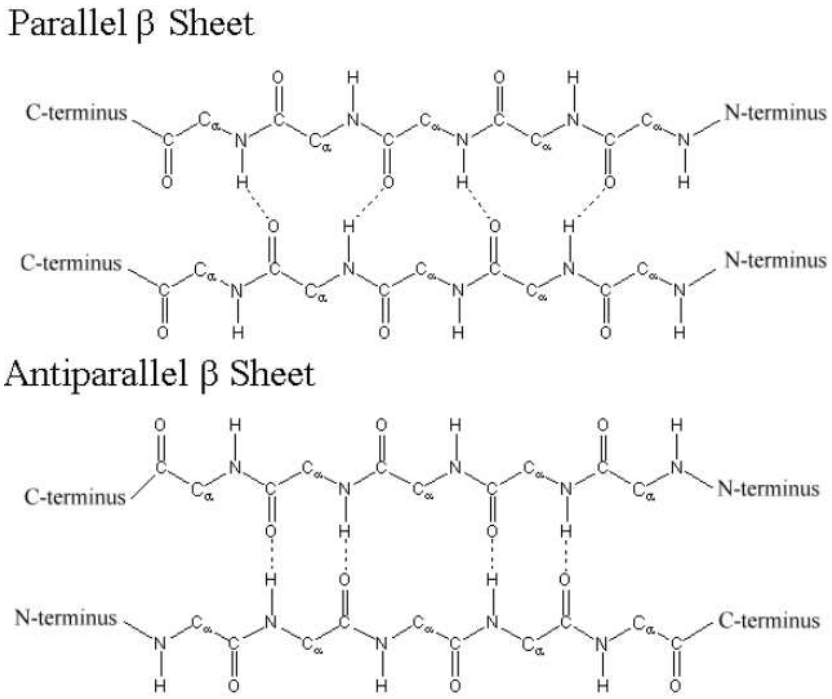

Figure 2.2: Formation of hydrogen bonds in the β-sheet

Protein motif and motif scanning methods

A protein motif is a short, conserved sequence of amino acids within a protein, often associated with a recognizable structural part that performs a specific function, such as binding, catalysis, or structural stability. These motifs provide insights into protein evolution since they are more conserved than other regions and often act as independent units within proteins. The presence of specific motifs helps predict the function of uncharacterized proteins.

Motif scanning methods are widely used techniques in bioinformatics for analyzing sequences. They aim to identify conserved patterns in protein sequences, detecting motifs similar to the query to determine family membership or structural and functional features. These methods take a motif, the query, as input and find similar subsequences within a given set of sequences, which is the chosen list of proteins. The output consists of sufficiently similar matches, referred to as the "response". Typically, these methods involve using a local alignment algorithm and a similarity function. The response is generated in two steps: first, all local alignment results are ranked by similarity to the query, and then only those above a certain similarity threshold are selected [10].

Chapter 3

Method - Outlines

3.1 A brief description

The goal is to improve the accuracy of motif scanning procedures by detecting as many significant motifs as possible (true positives) while minimizing the number of wrong assignments (false positives). Accuracy is measured by how closely the response matches the biologically relevant sequences in the sample. To enhance accuracy, we employ an approach based on pairwise similarity, examining not just the similarity to the query but also the mutual similarity among protein motifs within the response. Given a large response that presumably contains a significant number of false positives, we search for subsets where each pair of elements is sufficiently similar. The largest of these subsets is considered the new, modified response. This strategy is sensible because true positives are more likely to be similar to each other than false positives. We apply this approach to responses from two iterative motif scanners: PSI-BLAST and IGLOSS.

IGLOSS, which stands for Iterative Gapless Local Similarity Search, utilizes an input scale parameter, which is the confidence level, to set the threshold for what is considered "sufficient similarity." The response generated by IGLOSS consists of protein motifs with similarity greater than or equal to the specified scale. A higher scale parameter punishes deviations from the motif more severely, thus selecting more similar sequences. Consequently, the number of data points in the response is inversely proportional to the confidence level [11].

PSI-BLAST, or Position-Specific Iterative Basic Local Alignment Search Tool, differs from IGLOSS by allowing gaps. The confidence level in PSI-BLAST is the e-value, where a smaller e-value indicates greater similarity between the query and responses. With PSI-BLAST, a smaller e-value means a higher threshold, leading to the selection of more sim-

ilar sequences and resulting in fewer data points in the response. In contrast to IGLOSS, the number of data points in PSI-BLAST is proportional to the confidence level [1].

We systematically work with a lower threshold for two reasons: to hopefully avoid false negatives and to ensure we have a sizable set of positives to work with. Therefore, when using IGLOSS, smaller values are used for the input scale parameter, whereas with PSI-BLAST, larger values are used for the e-value.

3.2 Method summary

The analysis begins with a sequence of letters, known as the query, which possesses a specific property relevant to the study. Using defined criteria and techniques, we identify a relatively large set of similar sequences that potentially share this property with the query, referred to as positives. From this set, we aim to find a subset more likely to possess the same property as the query, known as true positives. Our goal is to identify a subset within the given set of positives, aiming to discard as many false positives as possible while retaining the true positives.

The developed method is based on analyzing pairwise similarity, unlike search engines, which calculate similarity to the query [10]. We assume that true positives are more densely clustered compared to false positives and will be located within a sphere. By finding this sphere, we obtain points such that each pair is "sufficiently" close to each other. However, we encounter two key challenges: the annotation for protein secondary structure is not reliable, and we lack a characteristic query. To tackle these challenges, we consider the "once green, always green" approach and the coherent query or relevant query approach.

The main goal of this method is to blindly identify a suitable subset without any prior knowledge of secondary structure annotation. By achieving this, we eliminate reliance on annotation accuracy, as successful identification of protein motifs sharing the same secondary structure annotation can occur without prior knowledge of their annotation. Thus, while the method initially depends on annotation accuracy for testing and validation, it aims to operate independently in practice.

3.3 Motivation

We closely follow the approach stated in [12], where a sphere predominantly containing true positives was identified blindly, i.e. without any prior knowledge of true positives. We have two images showing candidates for a specific family of plant enzymes. The first

image displays candidates for *Arabidopsis thaliana*, a wild plant. It shows a well-defined grouping that can be placed within a sphere, suggesting a single evolutionary origin. In this case, the criteria for the blind search is to find a sphere with the given radius containing the most elements. It turns out that the sphere satisfying this criteria, for *Arabidopsis thaliana*, predominantly consists of true positives. This demonstrates the method's effectiveness in identifying relevant candidates without prior knowledge of true positives.

In contrast, the second image shows candidates for *soybean*, where multiple groupings were identified that cannot be placed in a single sphere, possibly due to the hybridized nature of soybeans, indicating multiple evolutionary origins. Furthermore, a characteristic query is used to find the best candidates via a search engine. The characteristic query represents the characteristic features of the target. However, we do not have such a query and must ensure good candidates through other means.

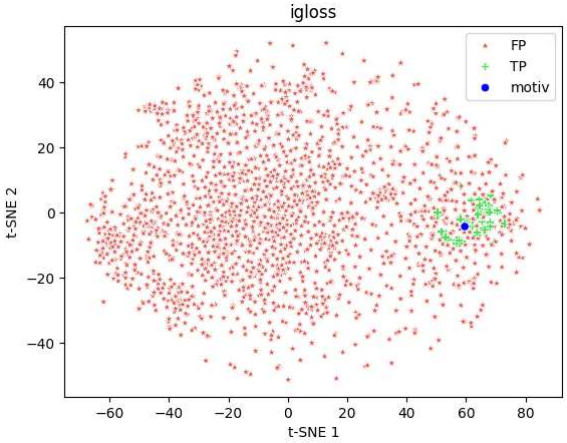

Figure 3.1: Arabidopsis thaliana

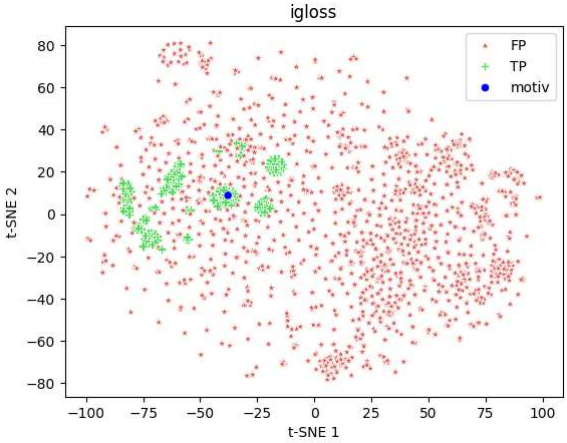

Figure 3.2: Soybean

Chapter 4

Method - Details

4.1 Procedure for identifying queries and good hits

We have a list of proteins with annotated secondary structure elements, internally named Structureome, consisting of 224,002 proteins. Protein secondary structure annotation describes the secondary structure of each amino acid subsequence. Here's a part of the first row in the list and its annotation:

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHL


----HHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHH-GGGGGG-TTTTT-
```

The procedure for identifying a query begins by selecting the secondary structure of interest, either Helix (H) or Sheet (S). We then extract sequences of amino acids of length $10 \leq L \leq 30$ where each element of its corresponding annotation is the letter H. From the provided example, potential queries are `EGEWQLVLHVWAKV` and `VAGHGQDILIRLFKS`. This process yields a list of all possible queries for helices, amounting to a total of 196,237. For sheets we extract sequences of length $8 \leq L \leq 15$ where each element of its corresponding annotation is the letter S.

A "good hit" denotes sequences with the same secondary structure annotation as the query. In other words, a good hit is simply another term for a true positive. Hits are the motifs that are "sufficiently" similar to the given query and are obtained from the search engine. For each hit, we examine its annotation. If the annotation consists only of H elements, allowing for one gap ("-"), it is considered a good hit, i.e., a true positive. For sheets, the process is analogous, with the symbol S used instead of H.

4.2 Data preparation

We perform sequence length adjustment to ensure uniformity in the length of retrieved sequences for analysis. Depending on the search engine used, IGLOSS maintains sequences at length L, while PSI-BLAST allows sequences to be shorter or equal to L. Shorter sequences in PSI-BLAST are extended with gaps ("-") to match the query length. For example, if the query is:

```
ENIKKEACWTISNIT
```

and the retrieved hit is:

```
DACWAISYLS
```

it is adjusted to:

```
ENIKKEACWTISNIT
-----DACWAISYLS
```

This adjustment ensures consistency in sequence length, facilitating accurate comparison and analysis in subsequent steps of the procedure.

Transition to Euclidean space

Sequences of amino acids are composed of letters, but the lack of a natural metric for comparing data creates an obstacle in conducting statistical analysis. To overcome this challenge, we represent amino acids with numerical values, enabling the transition to a vector space $\mathbb{R}^{L\times 5}$ [3]. Each amino acid is thereby defined by a 5-dimensional vector from the amino acid representation table 4.1. In this space, data is transformed into $L \times 5$-dimensional vectors.

Data standardization

To ensure searching for a favorable cluster is a reasonable strategy, data must first be standardized. If the variance of the data along one coordinate is significantly larger than along other coordinates, the Euclidean distance would be dominated by that coordinate, resulting in the loss of the spherical shape where all coordinates should have equal influence.

Standardization is adjusted to avoid division by very small numbers, thus reducing the possibility of outliers occurring. This adjustment is achieved through the formula:

$$x_i' = \frac{x_i - \bar{x}}{s + 0.1}$$

Amino Acid	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A	-0.591	-1.302	-0.733	1.570	-0.146
C	-1.343	0.465	-0.862	-1.020	-0.255
D	1.050	0.302	-3.656	-0.259	-3.242
E	1.357	-1.453	1.477	0.113	-0.837
F	-1.006	-0.590	1.891	-0.397	0.412
G	-0.384	1.652	1.330	1.045	2.064
H	0.336	-0.417	-1.673	-1.474	-0.078
I	-1.239	-0.547	2.131	0.393	0.816
K	1.831	-0.561	0.533	-0.277	1.648
L	-1.019	-0.987	-1.505	1.266	-0.912
M	-0.663	-1.524	2.219	-1.005	1.212
N	0.945	0.828	1.299	-0.169	0.933
P	0.189	2.081	-1.628	0.421	-1.392
Q	0.931	-0.179	-3.005	-0.503	-1.853
R	1.538	-0.055	1.502	0.440	2.897
S	-0.228	1.399	-4.760	0.670	-2.647
T	-0.032	0.326	2.213	0.908	1.313
V	-1.337	-0.279	-0.544	1.242	-1.262
W	-0.595	0.009	0.672	-2.128	-0.184
Y	0.260	0.830	3.097	-0.838	1.512
-	7.500	10.000	-6.000	-8.000	-1.000

Table 4.1: Amino acids and their corresponding factors

where $\bar{x}$ and s denote the arithmetic mean and standard deviation of the data, respectively. Each data point x_i is transformed accordingly.

4.3 Graphical representation of data

We utilize t-Distributed Stochastic Neighbor Embedding (t-SNE) to visually represent our data [9]. This statistical method effectively reduces the dimensionality of the data while preserving local structure. In other words, points that are close to each other in the original high-dimensional space remain close in the reduced-dimensional space. For example, if points B and C are both near point A in the high-dimensional space, and B is closer to A than C, t-SNE tries to maintain this local relationship in the reduced space so that B remains closer to A than C. By transitioning to two dimensions, it becomes possible to

visualize the separation of true positives from false positives.

4.4 Characteristic query

As mentioned earlier, one of the two challenges is the lack of a characteristic query. To address this, we begin by randomly selecting a query of length L and executing the entire procedure described thus far. Utilizing t-SNE, we monitor how the candidates behave spatially. If some clustering of true positives (TP) emerges, we extract motifs from this cluster to formulate a new query, internally named coherent or relevant. This coherent or relevant query then serves as our characteristic query. In cases where multiple clusters are identified, we prioritize the densest cluster for motif extraction [2]. And now, conducting the entire procedure with the established characteristic query results in much better clustering.

4.5 Confusion matrix - success rate

To measure the success of the method, we establish notation and define relevant accuracy measures. Subsequences/protein motifs annotated as having the same secondary structure are marked as condition positive (CP), while the rest are marked as condition negative (CN). Protein motifs returned by the method are denoted as positive (P), and the remaining as negative (N).

Depending on the actual and predicted state, each datum is assigned one of four outcomes: true positive (TP), false negative (FN), false positive (FP), or true negative (TN). True positives (TP) and false positives (FP) are defined as follows:

$$\mathrm{TP} = P \cap \mathrm{CP}$$

$$\mathrm{FP} = P \cap \mathrm{CN}$$

Likewise, true negatives (TN) and false negatives (FN) are defined as follows:

$$\mathrm{TN} = N \cap \mathrm{CN}$$

$$\mathrm{FN} = N \cap \mathrm{CP}$$

The relationships of these groups are illustrated with a confusion matrix, providing an overview of the classification result.

From the table, we can derive eight ratios, forming four complementary pairs, where the sum of each pair equals 1. These ratios are calculated by dividing the size of each of the four groups (TP, FN, FP, TN) by their sum in the corresponding row or column, representing the sizes of the remaining four groups (CP, CN, P, N).

		Predicted State	
		Positive (P)	Negative (N)
Actual State	Condition Positive (CP)	True Positive (TP)	False Negative (FN)
	Condition Negative (CN)	False Positive (FP)	True Negative (TN)

Table 4.2: Confusion matrix

Typically, the diagnostic ability of an application is assessed by comparing sensitivity or true positive rate (TPR):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{CP}}$$

and false positive rate (FPR):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{CN}}$$

Generally, when using a method like ours, there is an expected serious imbalance between the sizes of the condition positive (CP) and condition negative (CN) sets. CN is several orders of magnitude larger than CP, so for any reasonable test outcome, the false positive rate (FPR) will be close to 0. Consequently, precision or positive predictive value (PPV) is considered:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{P}}$$

In such cases, PPV and TPR are used as accuracy measures and are combined by their harmonic mean, called the F1-score:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$$

This is a standard way to measure the general accuracy of a search procedure by combining precision (PPV) and sensitivity (TPR). We could have applied this to our analysis provided that sets CP and CN were marked correctly. However, that is not the case.

4.6 Radius estimation

We aim to separate the true positives from the set of positives. The assumption is that true positives are close to each other and form a dense cluster in non-standardized data. After standardization, this dense cluster takes on a sphere-like shape. Therefore, it is reasonable to attempt to enclose such a cluster in a sphere [5].

The model

The goal is to estimate a sphere that separates true positives from false ones. We use numerical vectors of length $5L$ instead of amino acid sequences of length L. For intuitive understanding, we act as if we are working with sequences, not their numerical representations. We introduce the **conservation coefficient**, denoted by α, which represents how well the motifs in the sphere will be similar on average. Specifically, α represents the relative frequency of the dominant amino acid per column in a hypothetical motif profile, averaged over all columns.

The estimate

We define $\alpha \in (0, 1)$ as the conservation coefficient for any amino acid in the characteristic motif. The radius estimate relies on the parameter α, which is set a priori. To determine the radius, we calculate the expected distance between amino acid sequences sampled from the α-convex combination of distributions. This calculation assumes an average amino acid distribution along positions. Using a probabilistic geometry theorem, we can then estimate the radius. Finally, we adjust this estimate to account for standardized data.

Procedure

First, we need to define the average amino acid distribution R. The distribution R with probabilities for each amino acid is given by:

$$\mathrm{R} \sim \begin{pmatrix} A & R & N & D & C & Q & E & G & H & I & L & K & M & F & P & S & T & W & Y & V \\ 0.078 & 0.051 & 0.043 & 0.053 & 0.019 & 0.043 & 0.063 & 0.072 & 0.023 & 0.053 & 0.091 & 0.059 & 0.022 & 0.039 & 0.052 & 0.068 & 0.059 & 0.014 & 0.032 & 0.066 \end{pmatrix}$$

Here, r_i are the probabilities for amino acids, with $i \in \{1, 2, \ldots, 20\}$. Next, let A_i be the average distribution of amino acids with the conservation assumption of the i-th amino acid in the percentage $\alpha \cdot 100$. Specifically, A_i is given by:

$$A_i \sim \begin{pmatrix} a_1^i & a_2^i & \ldots, & a_{20}^i \\ p_1^i & p_2^i & & p_{20}^i \end{pmatrix}$$

where $i \in \{1, 2, \ldots, 20\}$.

The probability p_j^i is formulated as:

$$p_j^i = \alpha \cdot \mathbb{1}_{\{i=j\}} + (1 - \alpha) \cdot r_j, \quad j \in \{1, 2, \ldots, 20\}$$

where $\mathbb{1}_{\{i=j\}}$ represents the indicator function, which is equal to 1 if i equals j, and 0 otherwise. r_j are the probabilities from the average distribution R.

Consider two amino acid sequences of length L, sampled from a certain distribution. Let $X = (X_1, X_2, \ldots, X_L)$ and $Y = (Y_1, Y_2, \ldots, Y_L)$ represent the two observed sequences. We aim to calculate their expected distance. Using the definition of Euclidean distance and the linearity of expectation, we obtain:

$$\mathbb{E}\left[d^2(X, Y)\right] = \mathbb{E}\left[\sum_{i=1}^{L}(X_i - Y_i)^2\right] = \sum_{i=1}^{L}\mathbb{E}\left[(X_i - Y_i)^2\right]$$

Due to the lack of specific information about amino acids in any position, we assume they are some "average" amino acids, denoted as $\overline{X}$ and $\overline{Y}$. Therefore:

$$\mathbb{E}\left[d^2(X, Y)\right] = \sum_{i=1}^{L}\mathbb{E}\left[(\overline{X} - \overline{Y})^2\right] = L \cdot \mathbb{E}\left[(\overline{X} - \overline{Y})^2\right]$$

The expected square distance between two amino acids a_j^i and a_k^i sampled from distribution A_i is given by:

$$\mathbb{E}\left[(a_j^i - a_k^i)^2\right] = \sum_{j,k=1}^{20}(a_j^i - a_k^i)^2 p_j^i p_k^i$$

By averaging over the distribution R, we obtain the value of the expected square distance between two amino acids:

$$\mathbb{E}\left[(\overline{X} - \overline{Y})^2\right] = \sum_{i=1}^{20} r_i \sum_{j,k=1}^{20}(a_j^i - a_k^i)^2 p_j^i p_k^i$$

Consequently, we have:

$$\mathbb{E}\left[d^2(X, Y)\right] = L \cdot \mathbb{E}\left[(\overline{X} - \overline{Y})^2\right]$$

Theorem 4.6.1 (Jensen's Inequality)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function and $\mathbb{E}\left[\phi(X)\right] < \infty$. Then the following holds:*

$$\phi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\phi(X)\right].$$

Finally, using the Jensen's Inequality, we obtain an upper bound for the expected distance:

$$\mathbb{E}\left[d(X,Y)\right] \leq \sqrt{L \cdot \mathbb{E}\left[(\overline{X}-\overline{Y})^2\right]}$$

This upper bound serves as an approximation for the expected distance:

$$\mathbb{E}\left[d(X,Y)\right] \approx \sqrt{L \cdot \mathbb{E}\left[(\overline{X}-\overline{Y})^2\right]}$$

Theorem 4.6.2. *The expected distance between two points uniformly distributed in a sphere in an n-dimensional space approximates to $r\sqrt{2}$ as $n \to \infty$, where r is the radius of the sphere.*

By invoking the Theorem 4.6.2 [8], we obtain the estimate for the radius:

$$r_{\text{old}} = \frac{\mathbb{E}\left[d(X,Y)\right]}{\sqrt{2}} = \frac{\sqrt{L}}{\sqrt{2}} \cdot \sqrt{\mathbb{E}\left[(\overline{X}-\overline{Y})^2\right]}$$

After obtaining the estimate, we adjust it for standardized data. Let std_{old} and std_{new} be the standard deviations of data before and after standardization, respectively. Since radius and standard deviation are proportional quantities, the final radius estimate is given by:

$$r_{\text{new}} = r_{\text{old}} \cdot \frac{std_{\text{new}}}{std_{\text{old}}}$$

By substituting r_{old}, we obtain:

$$r_{\text{new}} = \frac{\sqrt{L} \cdot \sqrt{\mathbb{E}\left[(\overline{X}-\overline{Y})^2\right]}}{\sqrt{2}} \cdot \frac{std_{\text{new}}}{std_{\text{old}}}$$

As mentioned earlier, when using PSI-BLAST, we allow for gaps. Therefore, we need to incorporate gaps into the average amino acid distribution R. To adjust for the addition of a new element "-" (gap), we reduce the probability of all other elements (amino acids) by the same amount. The sum of the subtracted portion from each probability constitutes the probability of the added element "-" (chosen probability for a gap is 1%). This adjustment ensures that the total probability remains 1. The adjusted distribution R is then:

$$R \sim \begin{pmatrix} A & R & N & D & C & Q & E & G & H & I & L & K & M & F & P & S & T & W & Y & V & - \\ 0.0775 & 0.0505 & 0.0425 & 0.0525 & 0.0185 & 0.0425 & 0.0625 & 0.0715 & 0.0225 & 0.0525 & 0.0905 & 0.0585 & 0.0215 & 0.0385 & 0.0515 & 0.0675 & 0.0585 & 0.0135 & 0.0315 & 0.0655 & 0.0100 \end{pmatrix}$$

The rest of the procedure follows analogously.

Chapter 5

Results

For our analysis, it is necessary to have a relatively large set of positives. Generally, PSI-BLAST responses are small (around 100-200 positives), which tends to be insufficient for our analysis. Therefore, we also tried IGLOSS, whose response is more suitable for our analysis (around 4000-5000 positives). In addition to having a large set of positives, it is also very important to assume that each subsequence has a unique annotation. This assumption is crucial because it validates the concept of searching for secondary structure motifs.

5.1 An example

We start with a sequence of letters ENIKKEACWTISNIT. This is our query of length $L = 15$ and is identified as an alpha helix. Using the search engine PSI-BLAST, we try to identify a relatively large set of similar subsequences. However, due to PSI-BLAST's nature, some motifs in the response are shorter than the query length $L = 15$. To address this, we extend these shorter motifs with gaps to match the length of the query.

We represent amino acids using numerical values from the factor table 4.1. Protein motifs initially of length 15 are transformed into vectors of length 75. After standardizing the data, we visually observe the behavior of true positives and false positives using t-SNE. Each candidate is annotated, with true positives marked in green and false positives marked in red.

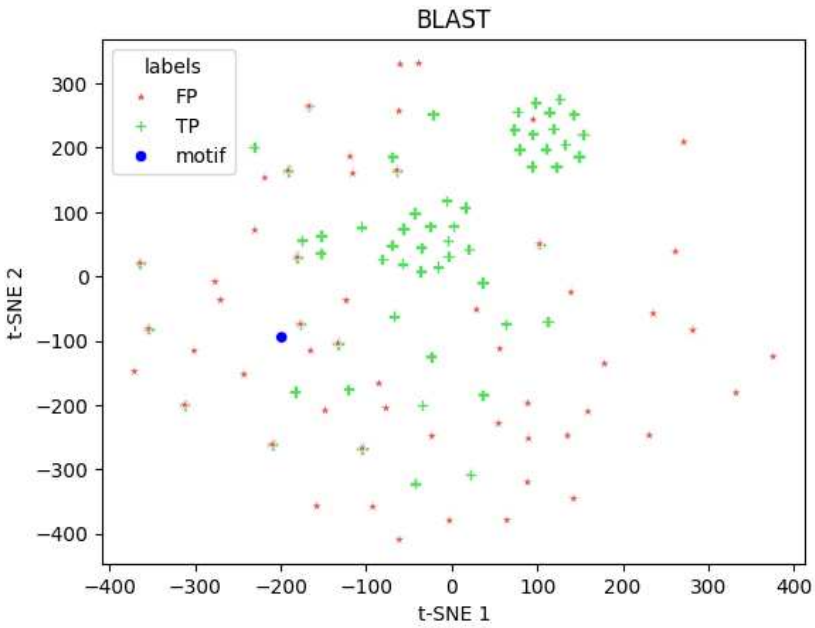

Figure 5.1: Some clustering

If some clustering of true positives (TP) emerges, we extract motifs from this cluster to formulate a new query, internally named coherent or relevant. This coherent or relevant query then serves as our characteristic query. In this case, where multiple clusters are identified, we prioritize the most dense cluster (cluster with most elements) for motif extraction.

The characteristic query is now a set of sequences of length $L = 15$ and it looks like this:

```
-DIKKEAAWAISNAT
ENIKKEACWTISNIT
EMLQLEAAWALTNI-
-SMLRNATWTLSN--
-NIQKEATWTMSNIT
EQILQEALWALSNI-
KSIKKEACWTISNIT
-SLIRTATWTLSNL-
--IQFESAWALTNI-
```

Conducting the entire procedure with the established characteristic query results in significantly improved clustering, as can be seen in Figure 5.2.

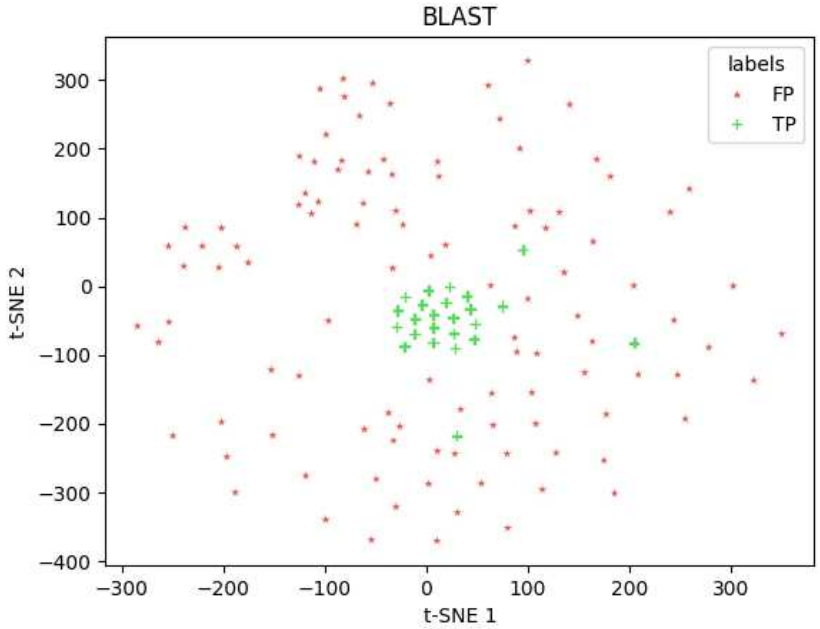

Figure 5.2: Better clustering

We attempt to place this improved cluster within a sphere that encloses nearly all of true positives. To define this sphere, we use the centroid of all true positives as the center and estimate the radius with $\alpha = 0.68$. We get:

$$\mathbb{E}\left[d(X,Y)\right] \approx \sqrt{L \cdot \mathbb{E}\left[(\overline{X} - \overline{Y})^2\right]} = \sqrt{15} \cdot 3.6799$$

$$r_{\text{old}} = \frac{\mathbb{E}\left[d(X,Y)\right]}{\sqrt{2}} = 10.0779$$

$$r_{\text{new}} = r_{\text{old}} \cdot \frac{std_{\text{new}}}{std_{\text{old}}} = 3.2453$$

With such a defined sphere, we successfully enclose almost all elements of the cluster, i.e., true positives. This can be seen in Figure 5.3.

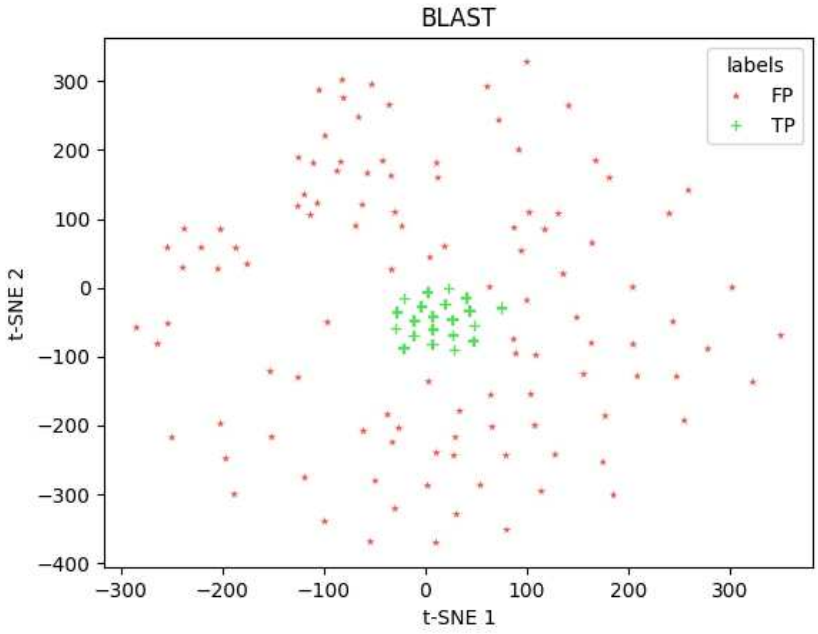

Figure 5.3: Sphere

In this case, the sphere contains only true positives.

We encountered a challenge in our analysis: obtaining a sufficiently large set of candidates (positives) was problematic with PSI-BLAST, prompting us to explore IGLOSS as an alternative. However, our attempt with IGLOSS response revealed significant annotation unreliability, as depicted in Figure 5.4.

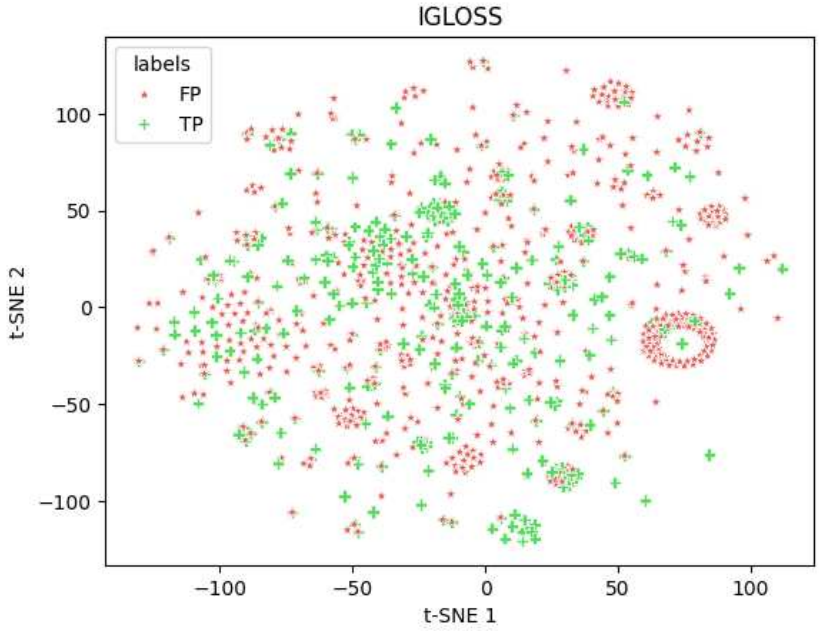

Figure 5.4: An example of "bad annotation"

The resulting chaos can be partially explained by the fact that all types of helices are marked the same. Locally, this chaos stems from the absence of a characteristic query and the lack of a unique (consistent) annotation of the subsequence, i.e., the same subsequence is sometimes marked as a helix and sometimes not in the list of proteins. We address this challenge with the "once green, always green" approach. This means if a subsequence is labeled as a helix once, we consider every such subsequence a helix regardless of its annotation. With these "corrected" data, we then take the centroid of all true positives as the center of the sphere and use the estimated radius where $\alpha = 0.68$ as the radius. For $\alpha = 0.68$, we calculate:

$$\mathbb{E}\left[d(X,Y)\right] \approx \sqrt{15 \cdot \mathbb{E}\left[(\overline{X} - \overline{Y})^2\right]} = \sqrt{15} \cdot 3.2973$$

$$r_{\text{old}} = \frac{\mathbb{E}\left[d(X,Y)\right]}{\sqrt{2}} = 9.0301$$

$$r_{\text{new}} = r_{\text{old}} \cdot \frac{std_{\text{new}}}{std_{\text{old}}} = 6.4756$$

Now, if a subsequence belongs to the sphere, we color every such subsequence green. Otherwise, it is colored red. This method successfully "corrected" the annotation, as can be seen in Figure 5.5.

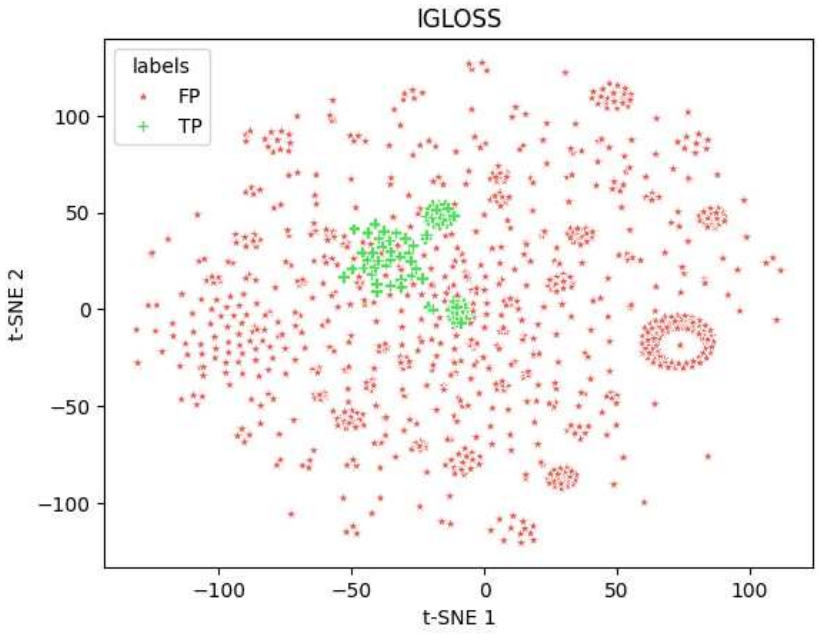

Figure 5.5: An example of "fixed annotation"

5.2 Queries for helices

Queries of length $L = 12$

For the query NGPLQWLDKVLT, the characteristic query is:

```
ASPEQWQEKAET
RKYLQWLTERLT
IHTLDWDDKMLE
RSALPTLKKVLT
YGILSHLDWVNN
ETPLQLLEKVKN
NKHLQDLMEGLT
HRALQLLDEVLH
MGSLYWLLPNLT
EGKLQHLENELT
SGWGQLLDRGAT
VGTLQRLPKVIG
NGPLQWLDKVLT
SKALDLLDKMLT
YGALRWFAGVLE
NGNLQYQGKDIT
```

which resulted in the following:

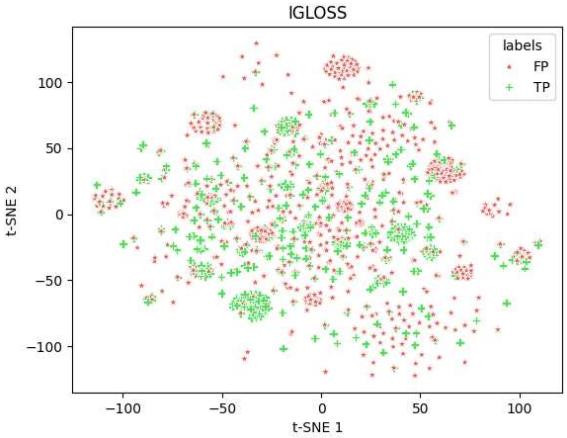

Figure 5.6: “Bad annotation”

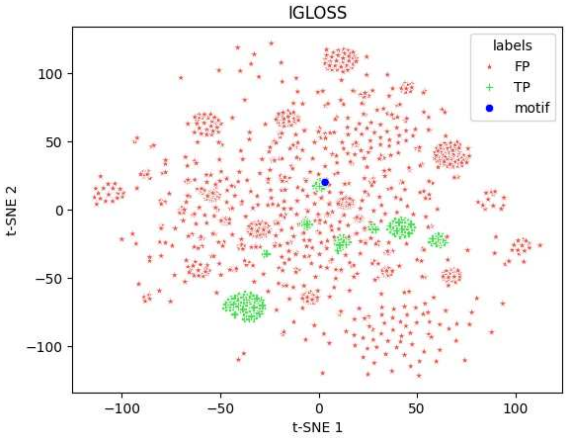

Figure 5.7: “Fixed annotation”

For the query TLVEQALKALGC, the characteristic query is:

```
WAVEQAHFALFF
PDVEQRFKAMGF
TGSMDALKAAGF
TLVEQALKALGC
SLVGQALFGDGA
ALQASALKAWGG
GLVFKWLKANGG
DATSATLKALGC
EVKEQAIWALGN
```

which resulted in the following:

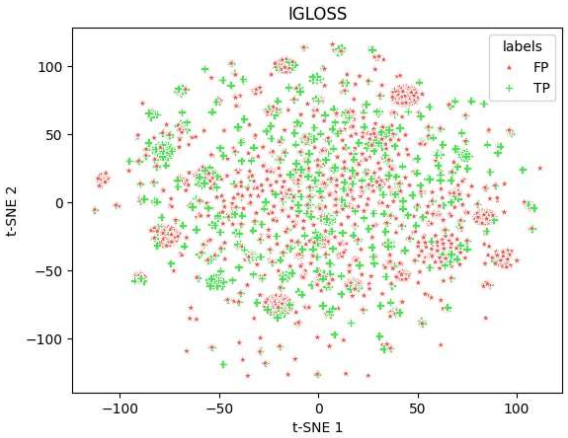

Figure 5.8: “Bad annotation”

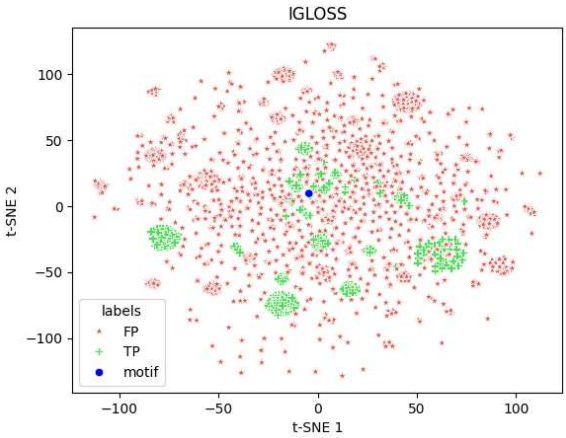

Figure 5.9: “Fixed annotation”

The query SYGLLGNSVDAL resulted in such good groupings that the coherent/relevant approach for the characteristic query was unnecessary.

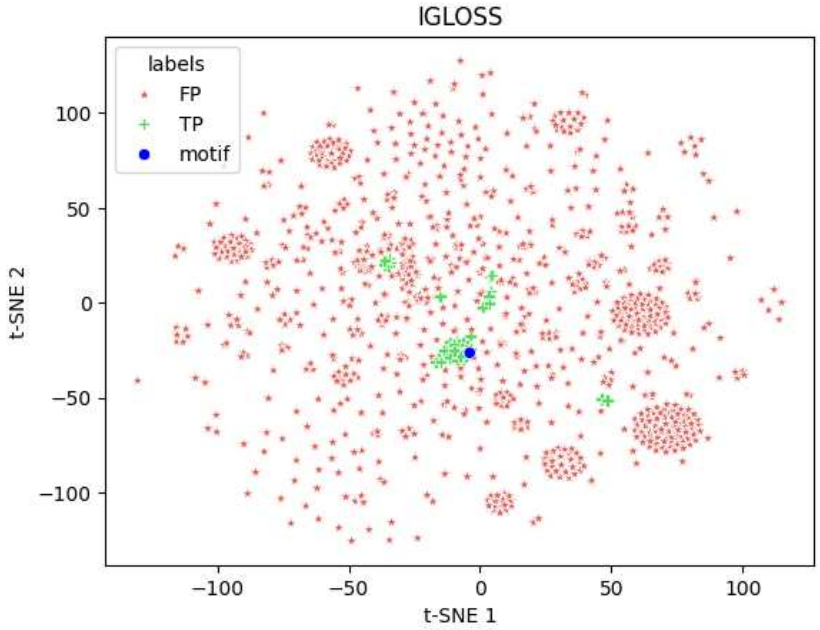

Figure 5.10: “Fixed annotation”

Query of length $L = 20$

For the query GAYRAMNKAALNFYETVRRD, the characteristic query is:

AFTLAVNVIAKKVTSTARID
GADCLMVKPAGAYLDIVREL
GTTCVTTGWGLTRYSTARID
NNYLNGLKLQGNFYNDAVID
GADMVMVKPGMPYLDIVRRV
GDFKAMYKALEGRPMTVRYL
GGFDSVNDWANKGYEVVVSN
EHLRTKNVAVRSFREGVRIT
ALVRTHSKKALMRYEDVYMP
MYLHKEQHSRLGFYSTARID
SFDRDKTIALIMNSSTARID
DAQGAMNKALELFRKDIAAK
GARRWINIDGKTMDITVKGL

GAGALAGAGALAGASTARID
INAGDLLKALLKPKSTARID
IQYLAVVASSHKGKSTARID
GADMLMVKPGMPYLDIVREV
KAYRRHDEVGTPFAVTVDYD
TELRLLTKALRPLPSTARID

which resulted in the following:

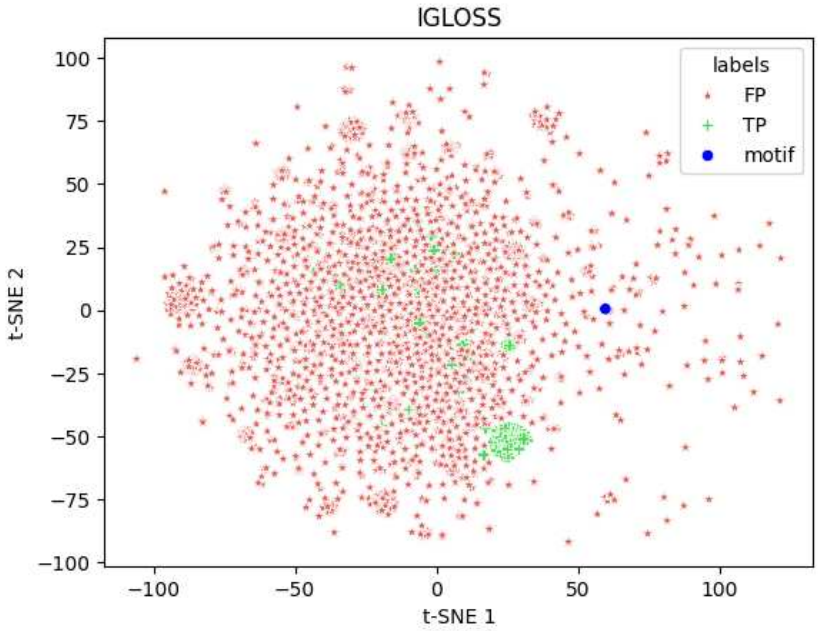

Figure 5.11: “Fixed annotation”

5.3 Queries for sheets

Queries of length $L = 8$

All the queries for sheets of length $L = 8$ resulted in such good groupings that the coherent/relevant approach for the characteristic query was unnecessary.
For the query SCHSGSCS the results were:

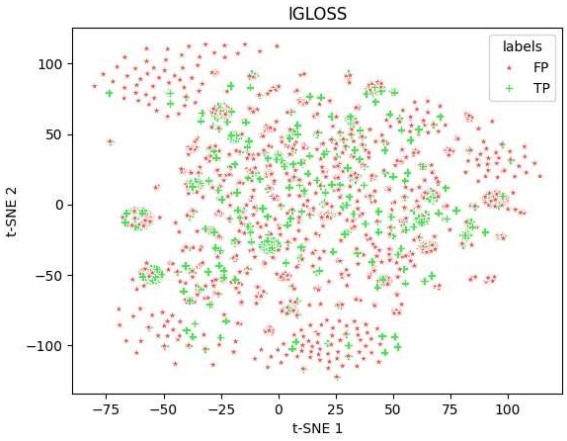

Figure 5.12: “Bad annotation”

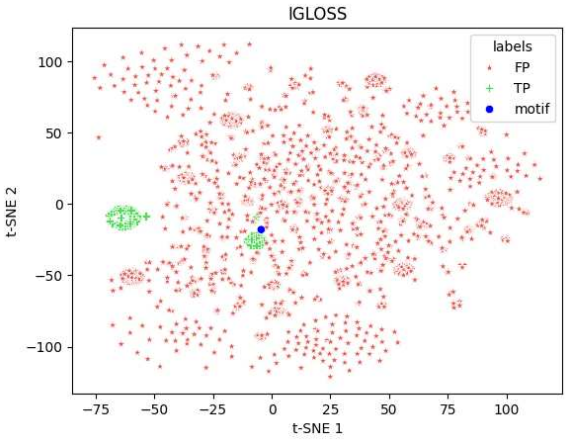

Figure 5.13: “Fixed annotation”

For the query SCQAGACS the results were:

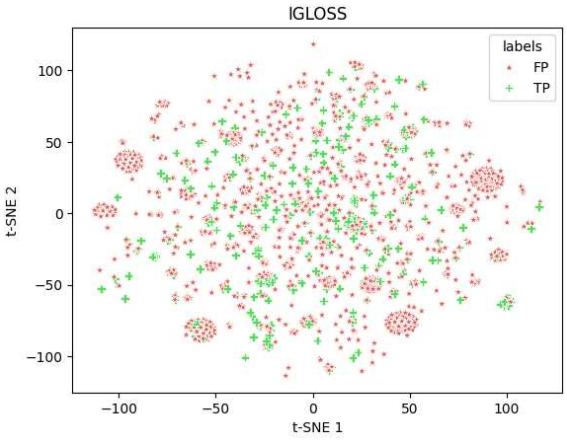

Figure 5.14: “Bad annotation”

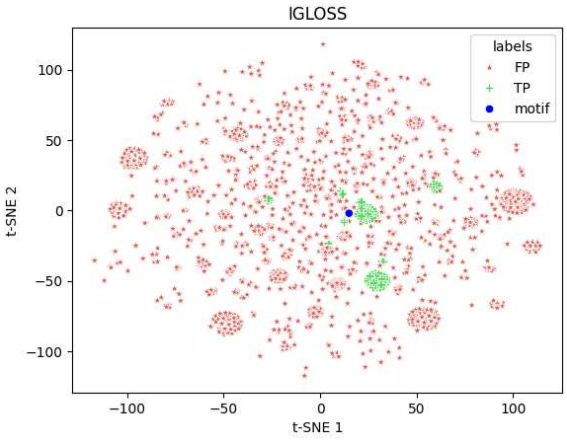

Figure 5.15: “Fixed annotation”

For the query GKDDYVKA the results were:

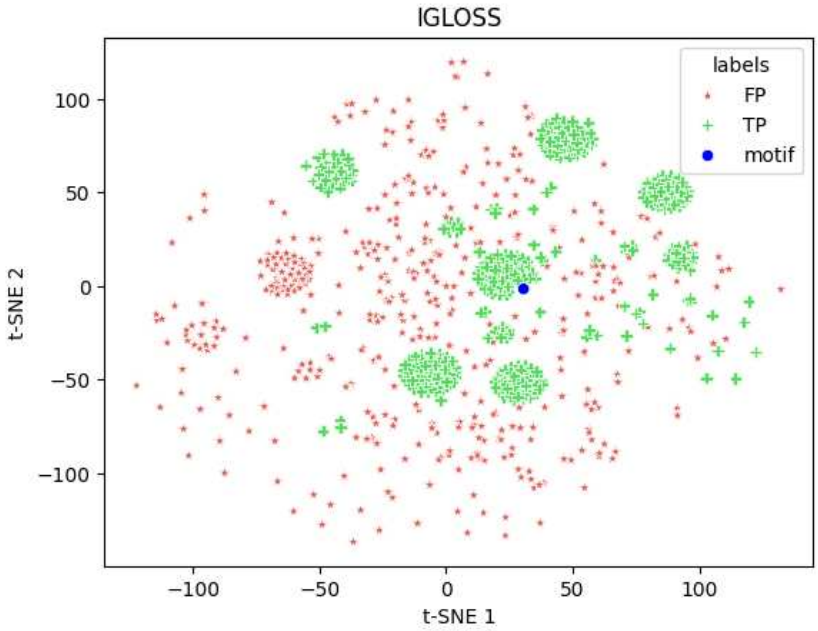

Figure 5.16: “Fixed annotation”

Queries of length $L = 9$

All the queries for sheets of length $L = 9$ resulted in such good groupings that the coherent/relevant approach for the characteristic query was unnecessary.

For the query DDYNTPDGT the results were:

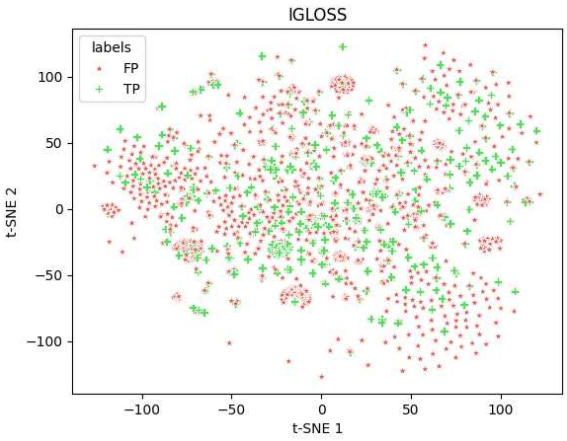

Figure 5.17: "Bad annotation"

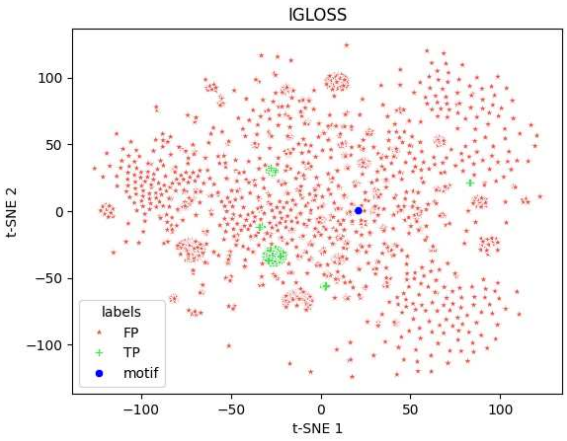

Figure 5.18: "Fixed annotation"

For the query NIDNEEIDE the results were:

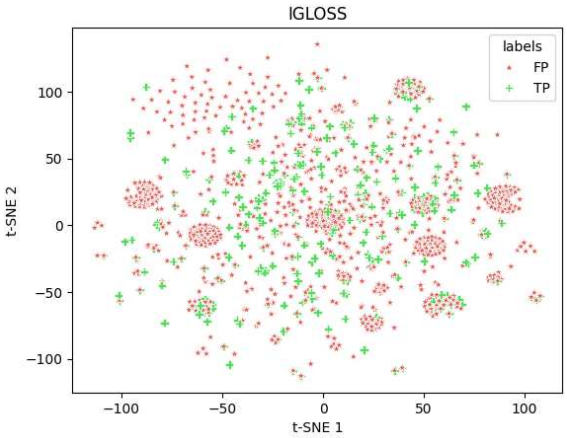

Figure 5.19: "Bad annotation"

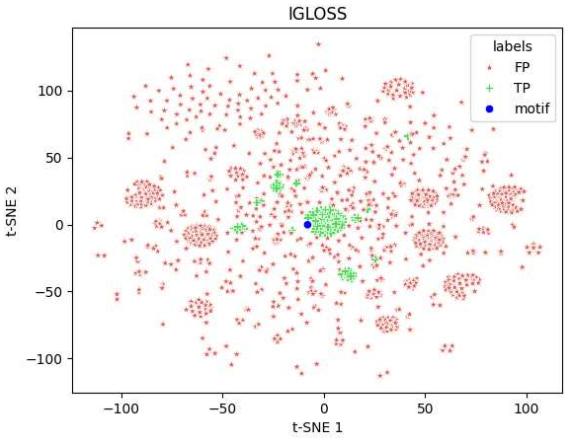

Figure 5.20: "Fixed annotation"

Query of length $L = 10$

The query WLRDFLWAQA resulted in such good groupings that the coherent/relevant approach for the characteristic query was unnecessary.

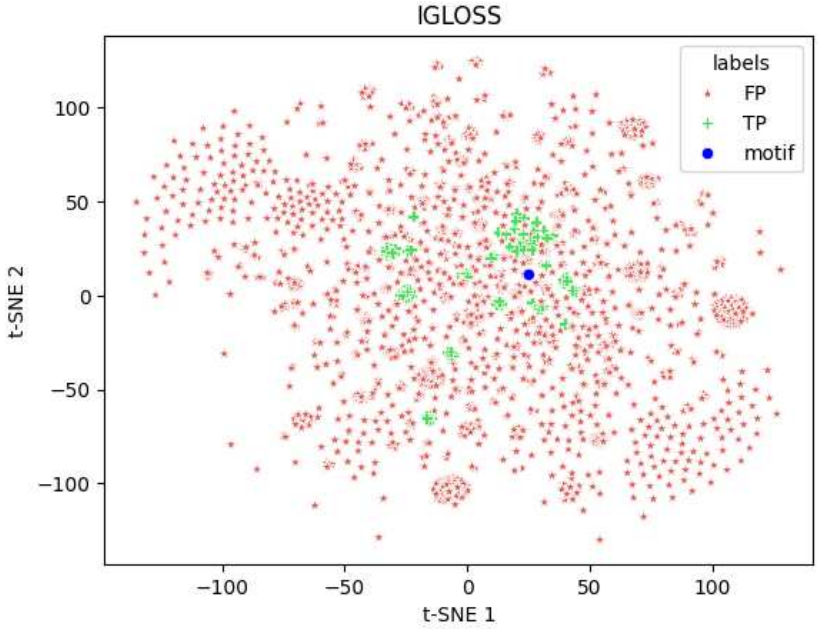

Figure 5.21: “Fixed annotation”

5.4 Conclusion

As we have seen with PSI-BLAST responses, gaps do not present a problem - they can be handled by the same techniques - in particular, standardization. Consequently, the issue of different motif lengths in PSI-BLAST response or other data sets containing gaps is successfully resolved. Additionally, the coherent/relevant query approach has proven to be a good option for addressing the lack of a characteristic query. By using the characteristic query obtained in this way, we have shown that secondary structure motifs cluster together, as can be seen from Figures 5.5, 5.7, 5.9, 5.10, 5.11, 5.13, 5.15, 5.16, 5.18, 5.20, 5.21. Typically, there are several clusters where good annotation would provide types. We have shown that with good annotation, it would be possible to attempt a blind search. In Figures 5.5, 5.7, 5.9, 5.13, 5.15, 5.16, 5.20, 5.21, we see that the structure of our response is similar to that of MADS-box family in soybean [12]; the clusters cannot be placed into a single sphere. Even if they could be placed into a sphere, as in Figures 5.10, 5.11, 5.18, it is not worth doing a blind search because we lack control due to poor annotation.

We are quite confident that we have improved the annotation, but we cannot be certain. To conclude, secondary structure motifs could be analyzed using an analogous technique to that used for plant enzymes or other motifs, but for that, we need more accurate annotation and a better characteristic query.

Bibliography

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res. **25** (1997), 3389–3402.

[2] Analytics Vidhya, *Introduction to k-means clustering algorithm*, currently available at `https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/`.

[3] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drüke, *Solving the protein sequence metric problem*, Proceedings of the National Academy of Sciences of the United States of America **102** (2005), 6395–6400.

[4] D. Bakić, *Linearna algebra*, Školska knjiga, Zagreb, 2008.

[5] D. Horvat, *Proteinski motivi i klasifikacija u proteinske familije*, Diploma thesis, University of Zagreb, Faculty of Science, 2023.

[6] M. Huzak, *Vjerojatnost i matematička statistika*, available at `http://aktuari.math.pmf.unizg.hr/docs/vms.pdf`, 2006.

[7] M. Iveković, *Traženje proteinskih motiva i klasifikacija*, Diploma thesis, University of Zagreb, Faculty of Science, 2022.

[8] M. G. Kendall and P. A. P. Moran, *Geometrical probability*, Hafner Publishing Company, London, 1963.

[9] M. Pathak, *Introduction to t-sne*, available at `https://www.datacamp.com/community/tutorials/introduction-t-sne`, 2018.

[10] B. Rabar, K. Nižetić, M. Zagorščak, K. Gruden, and P. Goldstein, *A clique-based method for improving motif scanning accuracy*, University of Zagreb, Faculty of Science, Mathematics Department and National Institute of Biology, Department of Biotechnology and Systems Biology.

[11] B. Rabar, M. Zagorščak, S. Ristov, M. Rosenzweig, and P. Goldstein, *Igloss: iterative gapeless local similarity search*, Bioinformatics **35** (2019), no. 18, 3491–3492, ISSN 1367-4803, https://academic.oup.com/bioinformatics/article/35/18/3491/5306940.

[12] J. Radnić, *Klasifikacija proteinskih fragmenata*, Diploma thesis, University of Zagreb, Faculty of Science, 2023.

[13] A. A. Schäffer J. Zhang Z. Zhang W. Miller S. F. Altschul, T. L. Madden and D. J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res. **25** (1997), 3389–3402.

[14] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[15] P. D. Sun, C. E. Foster, and J. C. Boyington, *Overview of protein structural and functional folds*, Curr Protoc Protein Sci **Chapter 17** (2004), no. 1, Unit 17.1.

[16] Wikipedia, *Protein secondary structure*, https://en.wikipedia.org/wiki/Protein_secondary_structure.

Summary

This thesis is concerned with the classification of protein motifs based on their secondary structure. A protein motif is a short sequence of amino acids that has remained partially conserved throughout evolution and can be associated with a recognizable part of the protein structure performing a distinct function. The goal is to improve the accuracy of motif scanning procedures by detecting as many significant motifs as possible (true positives) while minimizing the number of wrong assignments (false positives).

To enhance accuracy, we employ an approach based on pairwise similarity, examining not just the similarity to the query but also the mutual similarity among protein motifs. By describing amino acids using numerical vectors, the problem is placed in Euclidean space where distance is considered instead of similarity. The assumption is that protein motifs sharing the same secondary structure annotation will group together and be located within a sphere. By finding this sphere, we obtain points such that each pair is "sufficiently" close to one another.

However, we encounter two key challenges: the annotation for protein secondary structure is not reliable, and we lack a characteristic query. We addressed these challenges to the best of our ability, but due to unreliable annotation, we couldn't fully test the method. We showed that true positives do, in fact, cluster and could be found using this method. However, for this approach to be fully effective, we need more accurate annotation and a better characteristic query.

Curriculum Vitae

I was born in Zagreb on February 12, 1999. I began my education at Primary School Ivan Cankar in Zagreb, followed by attending IX. Gymnasium, also in Zagreb. After completing secondary education in 2017, I pursued the undergraduate degree in Mathematics Education at the Faculty of Science, University of Zagreb, which I completed in 2022. During that year, I enrolled in the graduate program in Biomedical Mathematics at the same faculty.