DETERMINING THE STATISTICAL SIGNIFICANCE OF EXTREME VALUES IN CLUSTERED DATA

An Undergraduate Research Scholars Thesis

by

ADVAIT PARULEKAR

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Thomas Ioerger

May 2018

Major: Computer Engineering

TABLE OF CONTENTS

ABSTRACT

Determining the Statistical Significance of Extreme Values in Clustered Data

Advait Parulekar
Department of Computer Science
Texas A&M University

Research Advisor: Dr. Thomas Ioerger
Department of Computer Science
Texas A&M University

A search through a large database for a match *similar* to the object being queried is commonplace. In the field of Bioinformatics, for example, this occurs during BLAST searches, in which an E-score is provided to reflect the significance of similarity of two gene sequences. In the case discussed in this work, the database consists of clusters (triplets) of amino acids found near the interface region of Protein Protein Interations (PPIs). The Exploring Key Orientations (EKO) algorithm needs to find similarity in structure between a peptidomimetic scaffold compound and a triplet present on such a PPI, and it is of interest to us to determine the triplets from within a large database of protein complexes that best fit the scaffold. It is our goal to determine when a "best match" thus acquired is statistically significant. We do this by parameterizing the space of triplets to find clusters, modeling a density distribution on the space, and fitting a Weibull distribution to determine a p value for a match. The inherently clustered nature of the triplet database affects the analysis of significance, and we propose a method to efficiently estimate the p value of a match score.

ACKNOWLEDGMENTS

I thank Dr. Thomas Ioerger for introducing me to this project. Besides suggesting the overall topic of research, he also suggested some of the ideas presented, and asked questions that led me to search for some of the answers here. If my work bears any resemblance of professionalism, it is due largely to him.

The development of the scripts related to EKO, most importantly the script which returns the most favorable conformation of each scaffold molecule to match to a PPI, and the extraction of the triplet database, was also done by Dr. Ioerger's group prior to my work.

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Proteins and Amino Acid Triplets

Proteins are highly variable biological molecules responsible for much of the physiology of life. Structurally, proteins are linear sequences of several (sometimes hundreds of) amino acids, which consist of a conserved peptide "backbone" and a variable "side chain" made up of amino acid residues. Protein function is highly dependent on their three dimensional spatial configuration, which is determined by the particular sequence of amino acids which constitute it. Protein structure comes from the lowest energy conformation of each of the variable bonds that make up the peptide backbone of the protein, as well as any variable bonds present in the side chains. Overall strucure is determined partially by backbone interactions, which yield the secondary structure, and side chain interactions. Side chains dictate protein structure by interacting with one another (for example, oppositely charged residues form electrostatic salt bridges) or with the aqueous solution they are in (it is entropically favorable to position hydrophobic residues such that they face other hydrophobic residues in the "interior", and it is energetically favorable for hydrophilic residues to be in contact with water, i.e. on the "outside" of the protein).

The carbon atom on the backbone that contains a bond to the branch point of the side chain is called the C_α atom and the carbon atom immediately adjacent to it on the side chain is the C_β atom. We define an amino acid triplet as being an ordered set of the C_α and C_β atoms of three (not necessarily consecutive, but clustered in space) amino acids of a protein. Figure 1.1 shows an example of such a triplet.

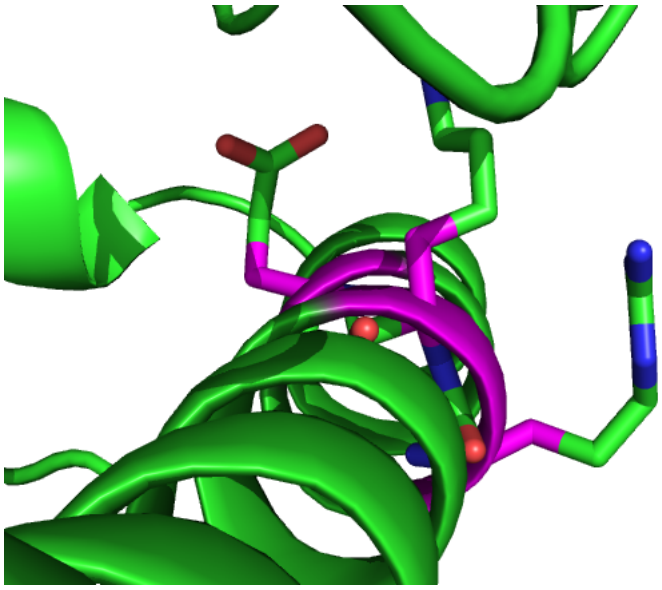
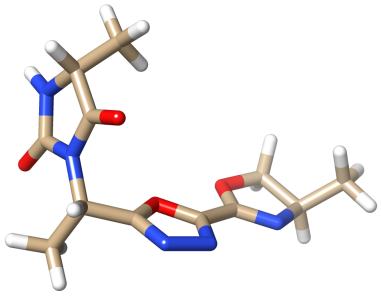


Figure 1.1: Example of a triplet on 1b0g (a Class 1 Histocompatibility Antigen). It is derived from the C_α, C_β bonds (shown in magenta) taken from an alpha helix.

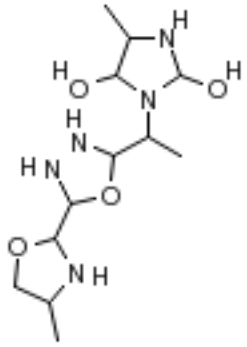
1.2 Peptidomimetics and PPIs

The immediate context of this work relates to the EKO algorithm which is used to design peptidomimetic compounds [1]. EKO searches (mines) for low RMSD matches of a scaffold molecule to interface regions from a large database of PPIs. The scaffold molecules are peptidomimetic compounds structured in such a way as to allow the attachments of residues to specific portions of the scaffold in an attempt to mimic a collection of residues (triplets) on a real protein. Potential scaffold molecules are screened based on their ability to mimic the conformation of the C_α and C_β atoms on some key amino acid triplet of the target protein. Figure 1.2 displays an example of a scaffold molecule.

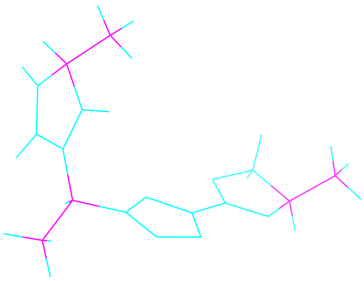
Molecules in solutions are dynamic due to the flexible nature of rotatable bonds. Quantum Molecular Dynamics (QMD) simulations of the scaffold moleucules, with methyl groups as place holders for the side chains, output a number of energetically feasible conformations for each of the stereoisomers of the scaffold. These conformations are filtered



(a) Example of a scaffold molecule



(b) 2D chemical structure of scaffold



(c) Magenta lines indicate the triplet that comes from this scaffold. The methyl groups on the triplets will be replaced by specific residues suited to the target interaction during synthesis

Figure 1.2

to exclude those conformations that match with a binding energy of greater than 3kcal/mol above the minimum across all conformations. The remaining conformations are clustered, and a representative is chosen from each of the clusters, which can then be "matched" against the set of triplets found in a target PPI. Match scores are the root mean square deviation between the scaffold C_α, and C_β atoms and their counterparts in the protein complex interface. The goal of this work is to place the relative match scores thus generated into perspective. Such an analysis can inform the screening of peptidomimetic molecules, the synthesis of which is a complex and expensive process. Peptidomimetic design in this work uses only the C_α and C_β atoms of triplets and the corresponding atoms on peptidomimetic scaffolds as an abstraction of the full structure.

Scaffold moleules begin with methyl groups as the C_β atoms in place of any larger side chains attached to the C_αs. To complete the analysis and simulation of a scaffold molecule, side chains are "stitched" on to the appropriate carbons. A semi-emperical force field can then be used to estimate the interaction energy ($\Delta\Delta G$) for the affinity with one member of the protein complex using AutoDock [2]. It is hypothesized that such mimics, once synthesized, can effectively perturb PPIs by competing with the protein they mimic, and methods have been developed to use advanced biomolecular software to model the interactions of proteins to develop such molecules [3]. Computational methods such as alanine scanning, which is used to gauge the effect of individual residues on binding affinity, and Generalized Born Surface Area (GBSA), which is used to model the interaction with the solvent, have been used to design peptidomimetics, for example, to inhibit the p53-MDM2 complex [4].

While the $\Delta\Delta G$ is a good approximation of the binding affinity of a scaffold, it is expensive to compute. We narrow our search using the RMSD matches. After all, in order for a scaffold to work, it must necessarily contain C_α and C_β atoms that align well with the mimicked protein.

1.3 Prior Work and Motivation

Any search of a database for similarity will result in a best match. In the context of EKO, these are the minimum match scores, that is, the lowest Root Mean Squared Deviation (RMSD) matches, of scaffold molecule conformations with triplets on a target PPI. It is our goal to help determine how significant such a match is. This relates to our concerns of specificity and off target effects [5]; we would like our scaffolds, and resulting peptidomimetics to match uniquely well. For instance, as we will see in Section 3.1 where we discuss secondary structure and clustering, if our scaffold mimics the structure of an alpha helix, we would expect to get very close matches to many interfaces, since triplets resemble alpha helices more often than average. While the overall binding affinity will depend largely on the particular amino acids that are stitched onto the scaffold, it is more likely that the scaffold will also perturb other PPIs if it comes from a geometric structure that is highly represented in the database.

A useful tool with which to approach this problem is the Extreme Value Distribution, which is defined as the distribution of extreme (in our case minimum) values of a random variable computed over several samples drawn from a target distribution (in our case the distribution of random triplet pair match scores). Given that the overall distribution of match scores conforms to certain analytical conditions, it is possible to model the extreme value distribution by a Weibull distribution [6]. This will be explored further in Section 2.5.

Baldi and Nasr conducted similar research in the context of chemical similarity [7]. They represent chemicals by an N bit bitstring $\mathbf{p} \in \{0,1\}^N$ using common binary fingerprints. Similarity was measured by the Tanimoto score, $\frac{\mathbf{p}_i \cap \mathbf{p}_j}{\mathbf{p}_i \cup \mathbf{p}_j}$. To model the Weibull distribution from a set of random Tanimoto match scores, the numerator and denominator of the score are modelled as correlated Normal random variables to the end of getting a p

value from a score. We ask the same questions they did. What should be threshold for a match to be considered a "good" match. For example, is an RMSD of $0.5\mathring{A}$ good enough? How can we interpret the significance of a match score conditioned on the query?

Such an analysis has also been conducted for the matching of triplets, but it assumed a uniform distribution of data. We will expand on this by taking advantage of the fact that clustering is to be expected due to the existence of triplets sourced from the relatively rigid secondary structures. This results in a non-uniform underlying density function, which we will try to account for.

1.4 PD-1/PD-L1 - a Motivating Example

Programmed cell death protein (PD-1) is a T-cell cell surface receptor that plays a role in downregulating the immune response and facilitating self-tolerance [8]. Figure 1.3 shows human PD-L1 in complex with murine PD-1. One of its two ligands, programmed death ligand 1 (PD-L1), which is expressed both on antigen-presenting cells and T-cells, is found to be upregulated in certain strains of cancer, and the interaction between the two has been studied as a possible target for drugs such as Durvalumab, Atezolizumab and Avelumab. Antibodies against PD-1 or PD-L1 have been shown to restore exhausted CD8 T cells during a chronic viral infection [9].

Such an interaction is a candidate for perturbation with a small molecule inhibitor. Peptidomimetic drugs structurally mimic a small part of a naturally occurring protein [10]. If the residues modeled by the peptidomimetic are critical for the PPI this can lead to a disruption of the interaction. In the case of PD-1/PD-L1, disrupting the association will lead to reduced self-tolerance. In the search of the appropriate inhibitory interaction, several scaffold molecules are tested, each of which concedes several conformations. When deciding between which scaffolds work best, the RMSD between the scaffold molecule and the corresponding residues on the partner protein is taken into consideration. However,

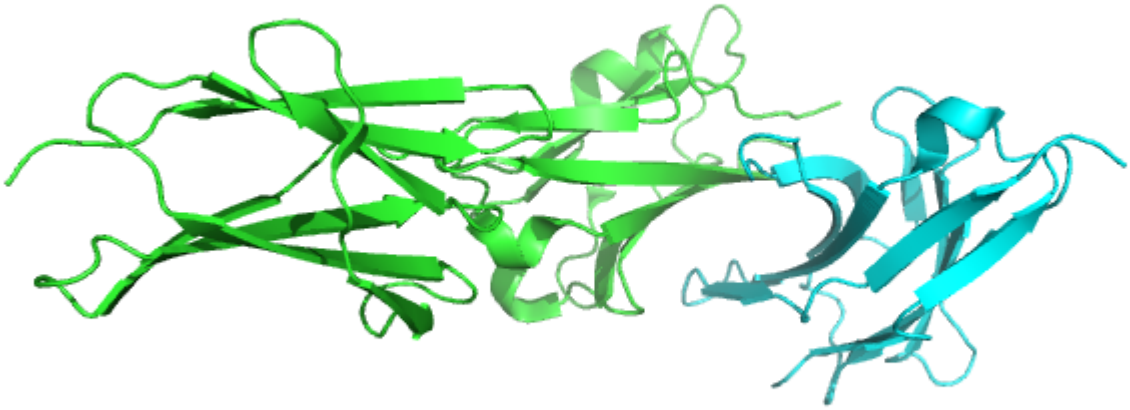


Figure 1.3: The PD1-PDL1 crystal structure (PDB: 3bik) with PDL1 (Programmed cell death 1 ligand 1) shown in green, and PD1 (Programmed cell death protein 1) shown in blue

in the interest of specificity, it may not necessarily be the scaffold that gives the optimal RMSD with a target triplet that is the ideal candidate, since the geometry of that scaffold may be such that it more regularly admits good matches in general. We also consider the conditional probability of finding a particular RMSD given the geometry of a scaffold molecule. This gives a statistical measure of the number of such matches across a large body of hypothesized protein complexes (all the proteins in the human body).

2. METHODS

2.1 RMSD Computation

Root mean square deviation (RMSD) among the 3 C_αs and C_βs is the primary metric between two triplets used in this work. Let a_i and b_i be a set of N points in $\mathbb{R}^n$, then we define the RMSD as being

$$d = \sqrt{\frac{1}{N} \sum_{1 \leq i \leq N} ||a_i - b_i||^2} = \sqrt{\frac{1}{N} \sum_{1 \leq i \leq N} \sum_{1 \leq j \leq n} (a_{ij} - b_{ij})^2} \tag{2.1}$$

Several methods to determine RMSD from spatial locations of each of the atoms of the triplets have been studied in the literature. To phrase the problem mathematically, given two sets of k points each $\mathbf{a}_i$ and $\mathbf{b}_i \in \mathbb{R}^n$, we must find the optimal rotation matrix, R, to minimize $J(R) = \sum_i |\mathbf{a}_i - R\mathbf{b}_i|$. A solution to find the optimal linear translation to minimize RMSD reported by Kabsch [11] uses singular value decomposition as follows. We first translate the triplets so that their centroids align. Let $A, B \in \mathbb{R}^{k \times n}$ be the matrices with rows $\mathbf{a}_i$, and $\mathbf{b}_i$ respectively. We compute the SVD of $AB^T = USV^T$ where U and V are orthogonal matrices and S is diagonal with entries in descending order. Our desired rotation matrix R (for the case $n = 3$) is given by:

$$R = U \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} V^T \tag{2.2}$$

where $d = \det(U) \det(V)$.

2.2 Parameterization

In order to avoid the difficulties in computing RMSD we parameterize the triplets using 9 parameters picked so as to capture geometric information. By embedding triplets into $\mathbb{R}^9$, we allow the construction of a well defined local density function, which can be used in the extreme value distribution (see Section 2.5). If we can arrange to use the Euclidean metric in $\mathbb{R}^9$ instead of RMSD, we can also save time in computing similarity between triplets. Nearest neighbours searches in Euclidean space can be computed in $O(\log n)$ rather than $O(n)$ in discrete space.

The parameters are chosen to be geometrically meaningful and permutation, translation and rotation invariant. Normalization (as described in Section 2.3) allows us to optimize the difference between true RMSD, and the RMSD estimate using euclidean distance. They are listed below.

The C_α (respectively C_β) triangle is the triangle formed by the three C_α (respectively C_β) atoms.

1. d_A, the sum of the distances between the C_α atoms (the perimeter of the C_α triangle).
2. v_A, the variance of side lengths in the C_α triangle.
3. γ_A, the variance of angle in the C_α triangle.
4. d_B, the sum of the distances between the C_β atoms (the perimeter of the C_β triangle).
5. v_B, the variance of side length in the C_β triangle
6. γ_B, the variance of angle in the C_β triangle.
7. τ, the average dihedral angle between the planes formed by
    - the C_{α_i}, centroid of the C_αs, centroid of the C_β

- centroid of the C_αs, centroid of the C_β, the C_{β_i}

over $i = 1, 2, 3$. This parameter is taken to suggest the average "torsion" of the triplet.

8. D, the distance between the C_α and C_β centroids

9. Γ, the angle between the plane formed by the C_α atoms and the segment connecting the centroids.

$v_A, \gamma_A, v_B, \gamma_B$ are chosen to depict the degree to which a triangle is scalene. Note that each of these parameters results from a symmetric function of angles and sidelengths. Together, these allow a triplet to be represented by a vector $\mathbf{p} = (d_A, v_A, \gamma_A, d_B, v_B, \gamma_B, \tau, D, \Gamma) \in \mathbb{R}^9$.

2.3 Scaling

In order to retain the RMSD between triplets in the form of Euclidean distance in the parameterized space, we use linear regression to determine the optimal weights to assign to each parameter. Let T denote a set of triplets, $d(\tau_1, \tau_2)$ represent the RMSD between triplet τ_1 and τ_2, and $\mathbf{p}_{\tau_1} \in \mathbb{R}^9$ denote the column vector representing the embedded triplet. Using numpy's linalg.lstsq, weights w_i are determined, one for each of the nine dimensions listed above, such that W, the diagonal matrix with entries w_i minimizes

$$\sum_{\tau_i, \tau_j \in T} (d(\tau_i, \tau_j)^2 - (\mathbf{p}_{\tau_i} - \mathbf{p}_{\tau_j})^T W (\mathbf{p}_{\tau_i} - \mathbf{p}_{\tau_j}))^2 \tag{2.3}$$

From here we can scale the parameter space; $\hat{\mathbf{p}}_{\tau_i} = W^{1/2}\mathbf{p}_{\tau_i}$ is used to represent the triplet, and we see that the euclidean distance estimate is $\hat{d}(\tau_i, \tau_j) = \sqrt{(\hat{\mathbf{p}}_{\tau_i} - \hat{\mathbf{p}}_{\tau_j})^T (\hat{\mathbf{p}}_{\tau_i} - \hat{\mathbf{p}}_{\tau_j})}$. See Figure 2.1 for an illustration of the agreement between scaled Euclidean distance and RMSD. The weights used are presented in Table 2.1.

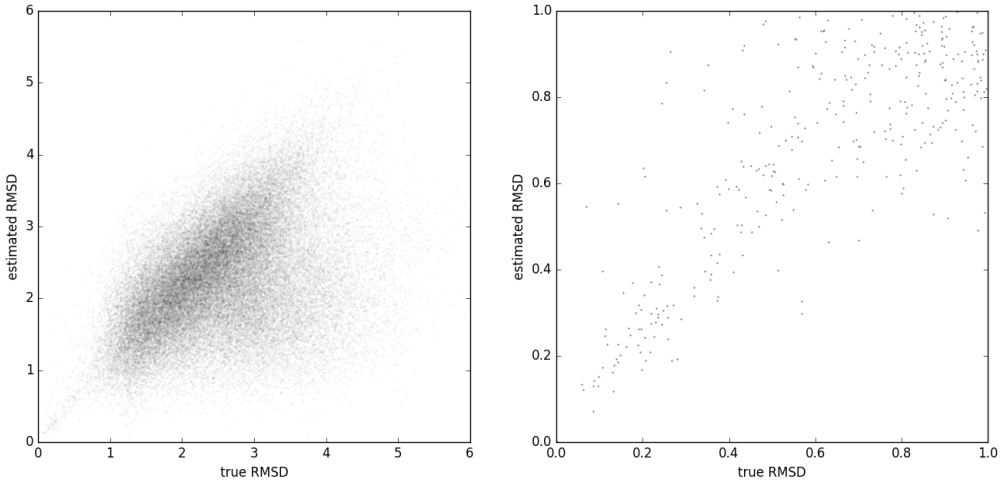


Figure 2.1: A comparison of weighted Euclidean distance between triplets on the x-axis and the true RMSD computed using the SVD decomposition described. The weights associated with each parameter are optimized so as to minimize the mean square difference between the two. On the right we have zoomed in to just plot the matches that are better than $1\mathring{A}$ RMSD.

Table 2.1: Weights used to scale the parameters so that the Euclidean distance between parameters optimally agrees with the true RMSD

d_A	v_A	γ_A	d_B	v_B	γ_B	τ	D	Γ
0.12	0.4	1.28	0.17	0.26	1.13	0.031	1.97	0.028

2.4 Clustering with DBSCAN

Density Based Scanning of Applications with Noise (DBSCAN) is used to cluster the data once it is embedded in $\mathbb{R}^9$ [12]. DBSCAN clusters data into regions of threshold densities. In summary, it categorizes points as core points, reachable points, and outliers using two user defined parameters: ϵ, indicating the radius of the spheres used in the algorithm and n, indicating the minimum number of other data points that need to be in a sphere of radius ϵ of a point for the point to be considered a core point. All points within

spheres of radius ϵ of a core point which are not core points are *reachable* and all others are outliers. DBSCAN suits the purposes of this project primarily because it is density based. As described in previous sections, much of our analysis is dependent on density, and so clusters in which points lie above a certain threshold of local density are useful to study to gain an understanding of the data. In particular, the premise of our work as mentioned in Section 1.3 is that certain triplets come from clusters of high density, and that the significance of those matches should be modulated based on an analysis of local density.

2.5 Extreme Value Distributions

To determine how statistically significant a match score is, we can compare it to the distribution of random match scores. A good match will give a score that corresponds to the tail region of the distribution.

Let random variables $X_1, X_2, \cdots X_n$ be independent and identically distributed (i.i.d.) from a cumulative distribution function (cdf) F. Let $m_n = \min X_1, X_2, \cdots X_n$ be the minimum (extreme value) of the X_i. The Extreme Types Theorem charaterizes the limiting cdf of m_n as n (the number of samples) grows large [13]. It states that such a limiting distribution can only be of three types. The type applicable to this scenario is the Weibull distribution, given by a cdf of the form $1 - e^{(\frac{x}{\lambda})^k}$ for some $\lambda, k > 0$.

Of interest to us is the distribution of the distance to the closest triplet from a given query point, and here we adapt the Weibull distribution for the purposes of closest matches in $\mathbb{R}^9$. Suppose the triplets are i.i.d according to some density function ρ in parameter space. Then the probability of any triplet being withing a distance of r of a point is proportional to the volume of the selected region. In this case we can just take it to be $\frac{Ar^d}{V}$, where d is the number of dimensions of the space, V is the total volume, and A is an appropriate constant. The probability that m_n (the distance to the closest triplet) is less

than r is $1-(1-\frac{Ar^d}{V})^n$, i.e. we have $P(m_n < r) = 1-(1-\frac{Ar^d}{V})^n$. From $(1-\frac{x}{n})^n \to e^{-x}$ and $V = \frac{n}{\rho}$, we get $P(m_n < r) = 1-(1-\frac{A\rho r^d}{n})^n \to 1-e^{-A\rho r^d}$. For $d = 9$, in $\mathbb{R}^9$, we get $A = \frac{32\pi^9}{945}$, which gives

$$P(m_n < r) =\to 1 - e^{-\frac{32\pi^9}{945}\rho r^d} \tag{2.4}$$

The results from a simulation consisting of 5000 points in $\mathbb{R}^9$ taken from a uniform distribution are provided in Figure 2.2 below, in which a histogram of nearest neighbour distances (best matches) is overlayed with the estimated extreme value distribution given above. We see that the theoretical distribution matches well with the simulation. A distribution of random distances from which the nearest distances come is also shown for perspective.

2.6 Local Density

Based on the previous section, we see that in order to make informed decisions about the significance of a match we need to know the local density in the parameter space around the point of interest. For each data point, we compute the number of neighbors within a ball of fixed radius. In order to do this efficiently, we use a kd-tree, which is shown to have $O(\log n)$ nearest neighbor search times [14]. This is where we take advantage of having embedded our data in $\mathbb{R}^9$. Such a calculation (nearest neighbours) on a discrete graph would have taken $O(n)$ time. It is also not clear how to define a local density function without the convenience of the Euclidean metric.

We can now model the density at all points in the parameter space using kernel density estimation.

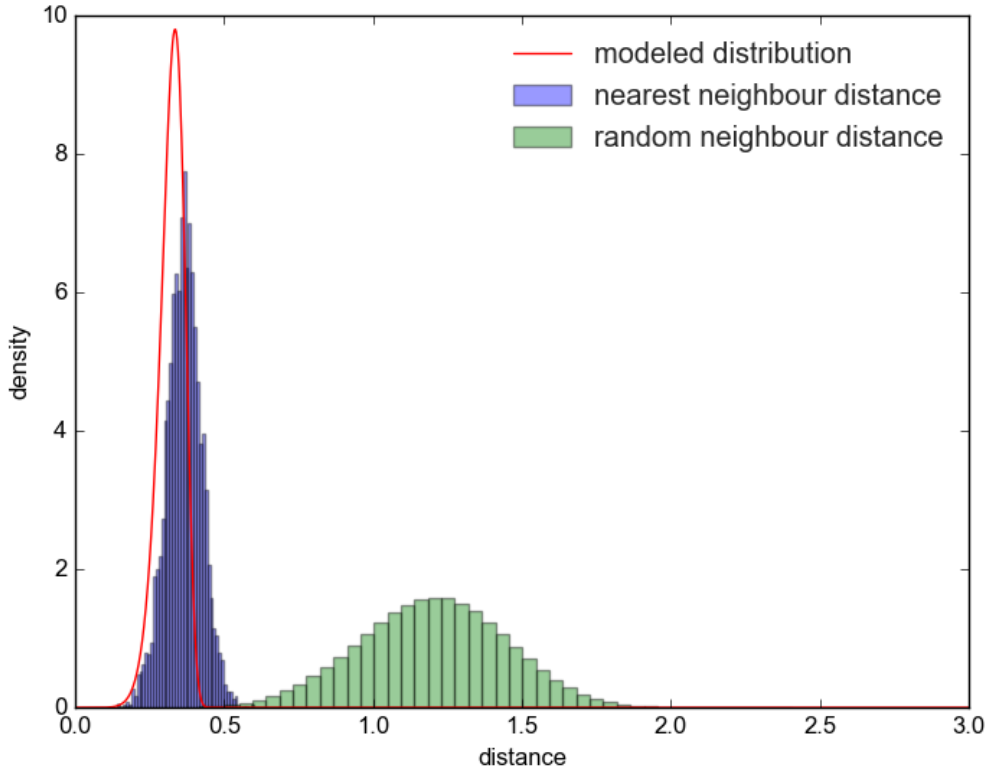


Figure 2.2: The histogram for the extreme value distribution from a simulation of random uniformly distributed data consisting of 10000 points in $\mathbb{R}^9$, plotted on top of the modeled distribution from the previous section.

2.6.1 Kernel Density Estimation

Kernel Density Estimation (KDE) is a non parametric way to model the underlying continuous density function associated with a data set as a "smoothed" function in order to estimate the local density at all points in space. Given a sample of independent and identically distributed data $(x_1, x_2 \cdots x_N)$, we consider the function $\rho(x) = \frac{1}{nh}\sum_i K(\frac{x-x_i}{h})$ for some smoothing parameter $h > 0$ and K is some function (kernel) that allows us to capture the concept of distance. This formulation allows us to compute the density ρ at any point x as an "average" (weighted by distance) of the densities at neighbouring

points in the dataset. The kernel K can be a variety of functions; we use the Gaussian: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. h is a smoothing parameter that is usually picked emperically.

The complete process is described in Figure 2.3.

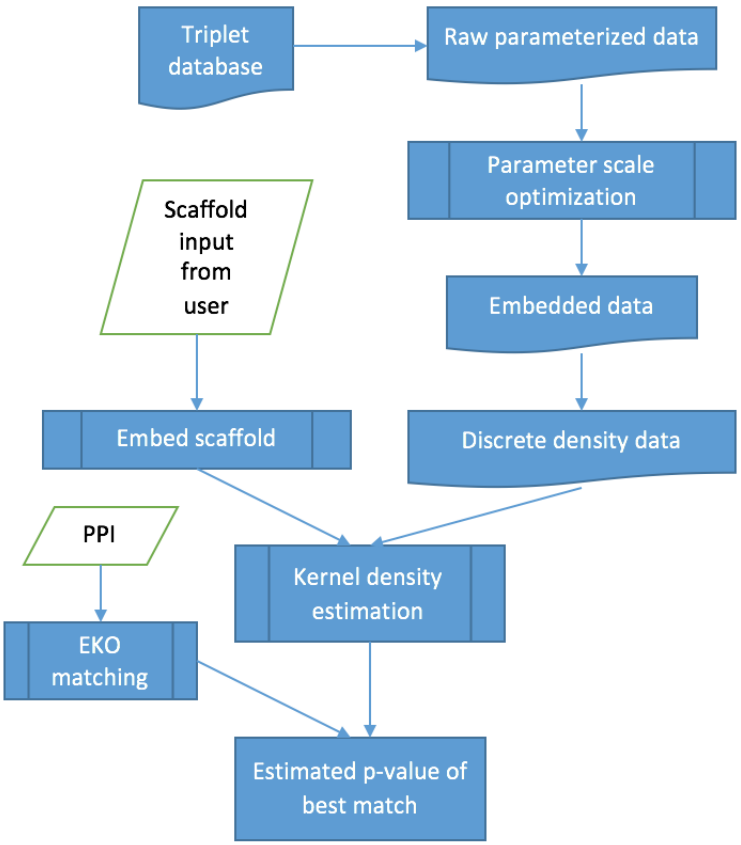


Figure 2.3: The complete process of using the triplet database in order to come to a understanding of the significance of a match.

3. RESULTS

In this work, we search a large database of approximately 5×10^6 triplets (defined in Section 1.1) chosen from $23,000$ proteins that yielded 2×10^5 heterodimer complexes to find low RMSD matches to some given chemical scaffolds. Triplets were filtered to include only those found at the interface region of the protein complex, where a sidechain is defined to be at the interface region if its C_α atom is within $4\mathring{A}$ of any atom on the other chain involved in the complex. The triplets were also filtered such that none of the distances between pairs of C_α atoms was more than $12\mathring{A}$ as these cannot easily be constructed on a scaffold.

Because a triplet is a representation of 6 atoms (3 C_αs and 3 C_βs), each of which lies in $\mathbb{R}^3$, the full triplet lies most directly in $\mathbb{R}^{18}$. Some of these dimensions are redundant, however, because a canonical triplet is invariant under rotations and translations. It can also be argued that since the $\overline{C_\alpha C_\beta}$ distances are roughly constant at $1.4\mathring{A}$, we can actually view them as lying in $\mathbb{R}^9$. For reasons described in Section 2.2, we embed the triplets into $\mathbb{R}^9$ using a parameterization scheme, the details of which are also presented in that section.

3.1 Clustering

After processing the data as indicated in Sections 2.2-2.3, a DBSCAN clustering was performed. The choice of parameters ϵ and n was determined so as to classify as large number of points into a minimum number of clusters. There are several such choices, and Figure 3.1 displays the results of the clustering using a pair ($\epsilon = 0.32, n = 30$) that gives 22 clusters. This choice results in 0.18% of the data being clustered, primarily into 6 large clusters. The DSSP database was used to compute the secondary structure for each residue in the clustered triplets. Results are shown in Figure 3.1.

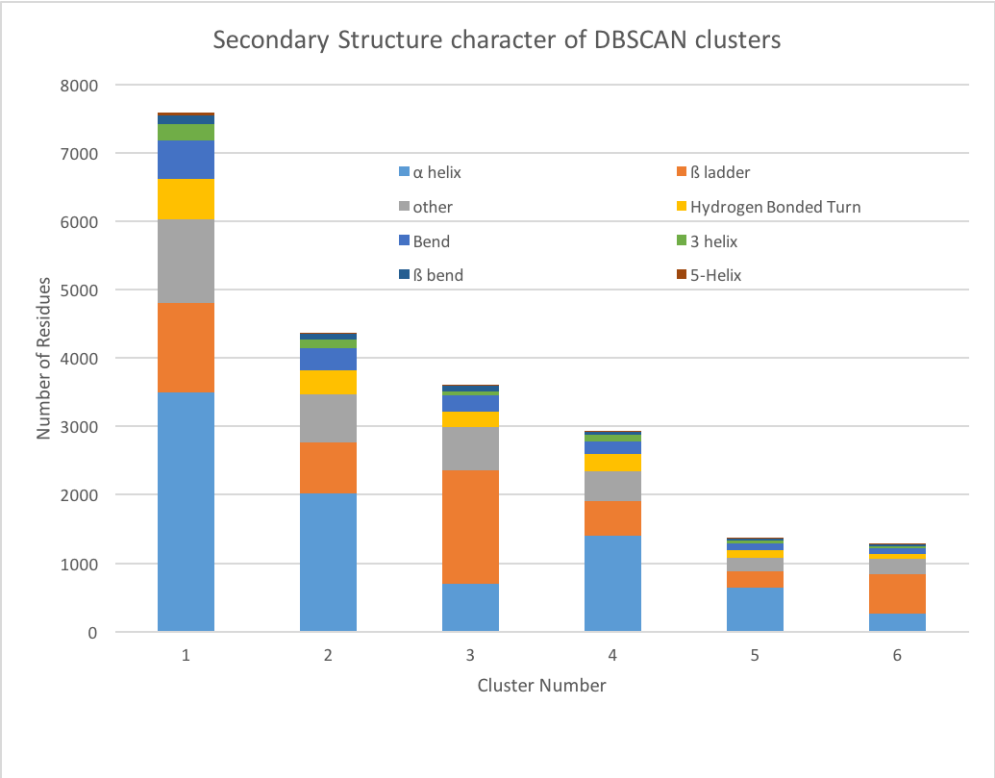


Figure 3.1: Cluster sizes for $\epsilon = 0.32$, $n = 30$ with standard scaling on raw parameters (pre-optimization). Most of the residues in each cluster come from α helices and β ladders.

3.2 RMSD Distributions and Extreme Values

We continue a discussion of Section 2.5 in the context of this work. The clusters returned by DBSCAN roughly identify regions with density above a threshold related to the parameters. Consider the distribution of general (random) RMSD scores $S_t(T) = \{d(\tilde{\tau}, \tau) | \tilde{\tau} \in t, \tau \in T\}$ for some sets t, T of triplets. We are interested in the distribution of $\bar{S}_t(T) = \{\min_{\tau \in T} d(\tilde{\tau}, \tau) | \tilde{\tau} \in t\}$. In order to highlight the dependence on local density of the resulting extreme value distribution, we present the distributions corresponding to the set t ranging over the various clusters found by DBSCAN in Figure 3.2. Upon comparing Figure 3.2 with the resulting extreme value distribution from t being a random sample of 1000 elements from the triplet database (in particular, not related to any cluster), shown in

Figure 3.3, we see that a blanket threshold for statistical significance is inappropriate. An RMSD of $0.25\mathring{A}$ is expected from the best RMSD for random matchings, but we would expect much a much better best match if it was known that one of the triplets in the match comes from a cluster.

Core points in DBSCAN are points with at least n neighbours within a distance of ϵ, meaning that the local density around each core point is at least $\frac{n}{\frac{32\pi^9}{945}\epsilon^d}$. The plot of the modeled Weibull pdf is overlayed on the histograms as an estimate of the actually extreme distribution. Note that the modeled distribution overestimates the minimum distances. This may be because the actual density of some of the points in the clusters may be much larger than the threshold required to be a part of the cluster.

3.3 EKO Matching

The PD1-PDL1 complex (PDB: 3bik) is used to evaluate the utility of this technique. We use a set of synthesizable scaffolds [15], and the set of interface triplets, both of which we represent in $\mathbb{R}^9$. The statistical significance of the lowest RMSD match from suitable (close to lowest energy) conformations of each of the stereoisomers of the scaffolds to each of the triplets in chain A of the complex interface is determined by computing the RMSD between the best conformer of the scaffold with every other triplet in the database (i.e., from other PPIs) and returning a true percentage of better RMSD matches. A density, computed using KDE, is used to generate a Weibull distribution for the best RMSD match scores. The RMSD of the closest match to any triplet on the PPI of interest is used to compute the probability of there being a better match among the remainder of the dataset, i.e., the p value of the match. The results are presented in the accompanying tables, Tables 3.1 to 3.4, along with images of the scaffold superimposed onto the best matching triplet on the target PPI, Figures 3.4 to 3.7.

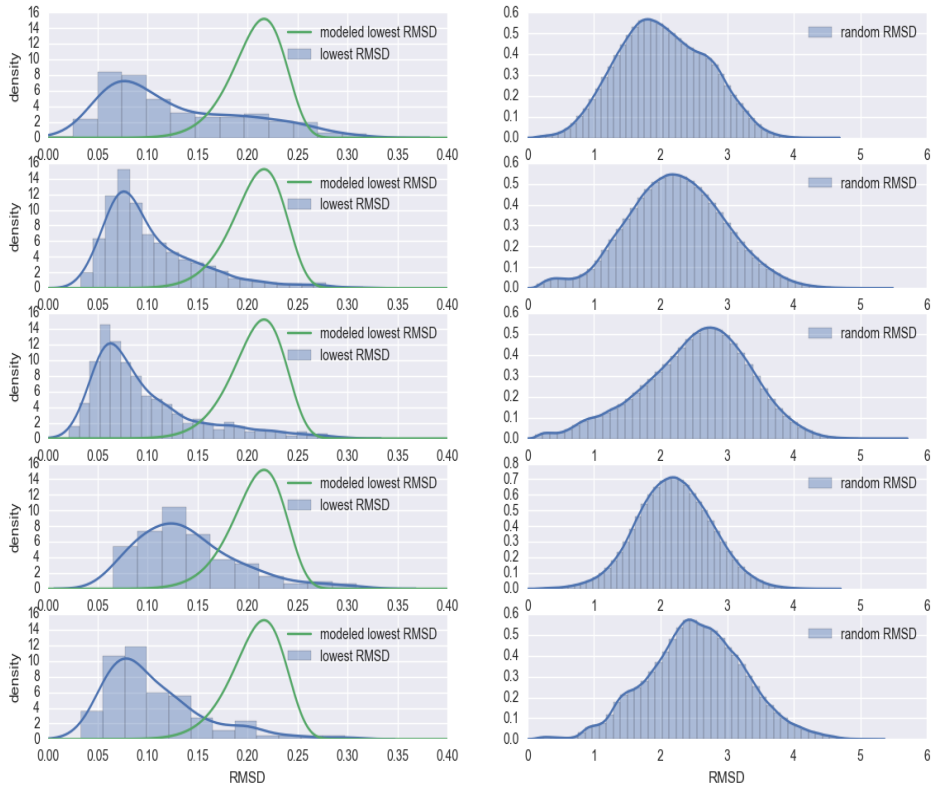


Figure 3.2: The distribution of best RMSD matches, i.e. $\bar{S}_t(T)$, (left) and general RMSD matches, i.e. $S_t(T)$, (right) for t ranging over the five largest clusters returned by DBSCAN are shown. T is a set consisting of 1% of the full triplet database.

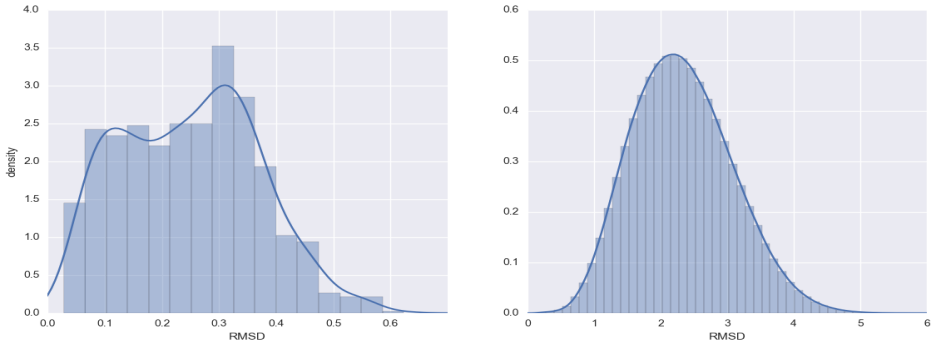


Figure 3.3: The distribution of best RMSD matches, i.e. $\bar{S}_t(T)$, (left) and general RMSD matches, i.e. $S_t(T)$, (right) where t is taken to be a set of 1000 random triplets, T is a set consisting of 1% of the full triplet database.

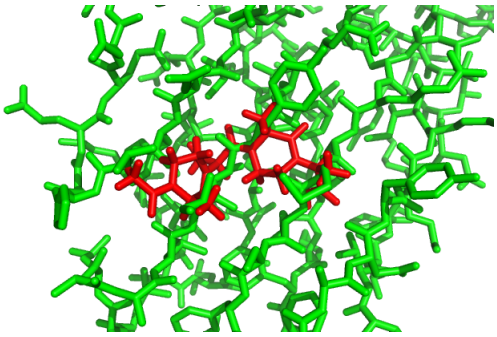


Figure 3.4: Chemotype 1: A scaffold shown superimposed onto the best matching triplet in the PPI.

Table 3.1: Chemotype 1: RMSD of best matches, percentage of better matches, i.e., the ratio of the number of triplets from the full database that match with a better RMSD to the total number of triplets, and the estimated probability of there being a better match in the database according to the density dependent distribution for each of the stereoisomers of the scaffold.

Chemotype 1, stereoisomer:	DLL	DLD	DDL	DDD	LLD	LLL	LDD	LDL
RMSD	0.48	0.6	0.67	0.46	0.44	0.46	0.51	0.52
better matches ($\times 10^{-5}$)	2.1	6.6	28	1.8	0.051	0.069	4.1	1.7
probability estimate ($\times 10^{-5}$)	0.56	1.4	19	0.23	0.12	0.17	0.74	0.08

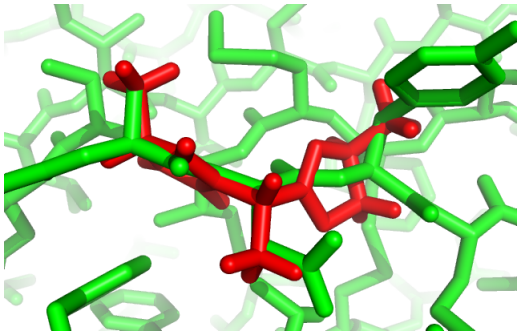


Figure 3.5: Chemotype 2

Table 3.2: Results for Chemotype 2

Chemotype 2, stereoisomer:	DLL	DLD	DDL	DDD	LLD	LLL	LDD	LDL
RMSD	0.45	0.35	0.51	0.56	0.38	0.53	0.46	0.35
better matches ($\times 10^{-5}$)	0.34	0.37	10.5	13.7	0.38	3.5	0.086	0
probability estimate ($\times 10^{-5}$)	0.22	0.028	0.44	0.98	0.045	0.96	0.32	0.011

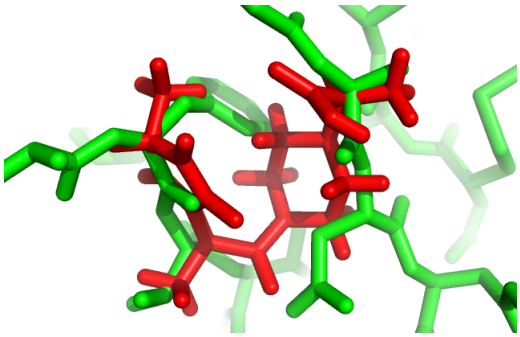


Figure 3.6: Chemotype 3

Table 3.3: Results for Chemotype 3

Chemotype 3, stereoisomer:	LDD	LDL	LLD	LLL	DDL	DDD	DLL	DLD
RMSD	0.58	0.59	0.45	0.49	0.49	0.46	0.63	0.49
better matches	0	0	0	0	0	0	0	0
probability estimate ($\times 10^{-11}$)	0	0.00014	0	0.54	2160	0.18	0.034	0.62

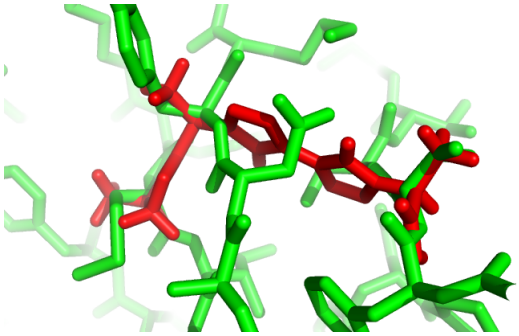


Figure 3.7: Chemotype 4

Table 3.4: Results for Chemotype 4

Chemotype 4, stereoisomer:	LDD	LDL	LLD	LLL	DDL	DDD	DLL	DLD
RMSD	0.52	0.44	0.33	0.29	0.44	0.48	0.4	0.45
better matches	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
probability estimate ($\times 10^{-5}$)	0.0044	0.22	0.02	0.0071	0.43	1.0	0.32	0.27

4. SUMMARY AND CONCLUSIONS

In this work, we presented a method to improve our understanding of the significance of a similarity match between two data points. The technique is particularly useful, as in this case, when only a subset of the data is visible in real time, and some heuristics from the global data which are computed offline are to be used. In the EKO matching algorithm, we start with scaffold molecules that can be synthesized, and a set of triplets from a protein interface, and look for the best match between any scaffold and any triplet therein. Comparing the match to the entire dataset is expensive. Instead, to quickly gauge the significance of the match, we use the local density to estimate how many better matches exist among all other protein interfaces. We may be interested in this just to filter match results, but there may also be more practical applications to do so. For example, the match score is fundamentally related to our estimate of the binding affinity of a peptidomimetic compound to a protein. In order to minimize the number of off target effects, we need *unique* matches from the triplet to the PPI.

4.1 Discussion of Results

4.1.1 Clustering

Whether or not a triplet in a DBSCAN cluster is a good first order measure of density. In Section 3.2, we study the RMSD distributions for triplets in clusters and find that if the clusters are too coarse, the threshold density is underestimated significantly, causing the distribution of expected best matches to be overestimated (Figure 3.2). A finer clustering, i.e. a DBSCAN performed with lower ϵ may improve the accuracy of this estimate.

4.1.2 Estimated probability of best match

We now attempt to answer the question we began with. A study of Tables 3.1 - 3.4 suggests that RMSD matches alone cannot be used as measures of statistical significance. Indeed, if we compare chemotype 1, stereoisomer DLD with chemotype 3, stereoisomer LLL, we see that chemotype 1 matches with much better RMSD ($0.35\mathring{A}$ against $0.49\mathring{A}$). However, a thorough search through the triplet database reveals that while there are 20 better matches in the database for chemotype 1, there are none for chemotype 3. Moreover, our probability estimate agrees; the probability estimate of the best match being better for chemotype 1 is 2.8×10^{-6} against 5.4×10^{-12}. We see that while the computed density based probability estimate presented is not an accurate measure of the number of better matches, a comparison of probabilities may lead to meaningful conclusions about the level of uniqueness of a match. The density based estimate has the benefit of calculation time, in that it runs much faster than a full database search. With large datasets, if an estimate is needed in real time, our method may be more practical to use.

4.2 Improvements and Future Work

The specific geometric parameters were chosen subjectively in an attempt to capture the essence of a triplet. It is likely that different parameters, chosen in a more objective manner, could lead to improved agreement between the RMSD estimate and true RMSD. What is desired is just a canonical representation of a set of three vectors.

In performing Kernel Density Estimation we use a similar bandwidth as was chosen for ϵ in the DBSCAN. Different bandwidths lead to different density esimates, and going forward, there may be better ways to decide on one.

A limitation of our work is that the scaffold conformation is picked such that the match with the best triplet within the target PPI is optimized, and the optimized conformation of the scaffold is then compared with the rest of the database, i.e., to triplets from other

PPIs. This is a bias and results in the extremely good matches (high uniqueness) matches reported. A complementary analysis which uses the scaffold before optimization is also needed.

This technique has a wide range of applications. For instance, given a set of people a social network, we might like to know which pair has a high probability of becoming friends. Rather than focusing exclusively on the similarities in opinion between people, we might also like to know, for each opinion on each subject, some measure of the "peculiarity" of that opinion. Two people who share somewhat their stance on a peculiar set of opinions, are more likely to become friends than two people who match more exactly but support a set of widely held beliefs. This satisfies the constraints of our problem - we are looking to place similarities generated from within a subset of a population into the context of global similarities. As we have seen, doing so by estimating the local density function and modeling the Weibull distribution may lead to good real time estimates.

REFERENCES

[1] E. Ko, A. Raghuraman, L. M. Perez, T. R. Ioerger, and K. Burgess, "Peptidomimetics, a synthetic tool of drug discovery," *Journal of the American Chemical Society*, vol. 135, pp. 167–173, 2009.

[2] G. Morris, R. Huey, W. Lindstrom, M. Sanner, R. Belew, D. Goodsell, and A. Olson, "Autodock4 and autodocktools4: automated docking with selective receptor flexiblity," *J. Computational Chemistry*, 2009.

[3] A. M. Watkins, R. Bonneau, and P. S. Arora, "Modeling and design of peptidomimetics to modulate protein–protein interactions," pp. 291–307, 2017.

[4] C. H. Zhong H, "Computational studies and peptidomimetic design for the human p53-mdm2 complex," *Proteins*, vol. 58(1), pp. 222–34, 2005.

[5] Y. ER, C. AE, and J. AN, "Prediction of off-target effects through data fusion," *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, vol. 19, pp. 160–171, 2014.

[6] S. S. Weibull W, "A statistical distribution function of wide applicability," *ASME Journal of Applied Mechanics*, pp. 293–297, 1951.

[7] P. Baldi and R. Nasr, "When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values," *Journal of Chemical Information and Modeling*, vol. 50, no. 7, pp. 1205–1222, 2010. PMID: 20540577.

[8] D. Y.-w. Lin, Y. Tanaka, M. Iwasaki, A. G. Gittis, H.-P. Su, B. Mikami, T. Okazaki, T. Honjo, N. Minato, and D. N. Garboczi, "The pd-1/pd-l1 complex resembles the antigen-binding fv domains of antibodies and t cell receptors," *Proceedings of the National Academy of Sciences*, vol. 105, no. 8, pp. 3011–3016, 2008.

[9] D. L. Barber, E. J. Wherry, D. Masopust, B. Zhu, J. P. Allison, A. H. Sharpe, G. J. Freeman, and R. Ahmed, "Restoring function in exhausted cd8 t cells during chronic viral infection," *Nature*, vol. 439, pp. 682–683, 12 2005.

[10] J. Vagner, H. Qu, and V. J. Hruby, "Peptidomimetics, a synthetic tool of drug discovery," *Current Opinion in Chemical Biology*, vol. 12, no. 3, pp. 292 – 296, 2008.

[11] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, 1976.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," pp. 226–231, 1996.

[13] T. L. Fisher RA, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 180–190, 1928.

[14] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, pp. 509–517, Sept. 1975.

[15] J. Taechalertpaisarn, B. Zhao, X. Liang, and K. Burgess, "Small molecule inhibitors of the pcsk9-ldlr interaction," *Journal of the American Chemical Society*, vol. 140, pp. 3242–3249, 03 2018.