Birkbeck Biology Department

**Protein Geometry and**

**Divergence from Trivial**

|  |
| --- |
| Student: Rachel Alcraft  Supervisor: Dr Mark A. Williams  Date: 21/02/2022 |

Table of Contents

[Figures, Tables, Equations ii](#_Toc96583521)

[Introduction 1](#_Toc96583522)

[1 Williams Divergence from Trivial 3](#_Toc96583523)

[1.1 Example for no variance 3](#_Toc96583524)

[1.2 Trivial examples 5](#_Toc96583525)

[1.3 Introducing variance 6](#_Toc96583526)

[1.4 Normalisation of the divergence 8](#_Toc96583527)

[1.5 Divergence p-values 9](#_Toc96583528)

[1.6 Results 10](#_Toc96583529)

[2. Idealised Protein Geometry 11](#_Toc96583530)

[2.1 Creating a synthetic model 11](#_Toc96583531)

[2.2 Idealised plots Comparison 11](#_Toc96583532)

[2.3 Results 11](#_Toc96583533)

[Discussion 11](#_Toc96583534)

[Future Direction 11](#_Toc96583535)

[References 11](#_Toc96583536)

# Figures, Tables, Equations

[Figure 1 highly correlated geometric plots. 1](#_Toc96583537)

[Figure 2 Plots for linear example 3](#_Toc96583538)

[Table 1 Varying parameters for the divergence 4](#_Toc96583539)

[Equation 1 Formula for the divergence with no variance 4](#_Toc96583540)

[Figure 3 Trivial relationships 5](#_Toc96583541)

[Figure 4 Increasing sample size smooths bins 6](#_Toc96583542)

[Figure 5 Sample size improves divergence metric 7](#_Toc96583543)

[Equation 2 The Williams Divergence from Trivial metric 8](#_Toc96583544)

[Figure 6 The p-value calculated for 3 distributions 9](#_Toc96583545)

[Figure 7 Top diveregnces calculated for high resolution glycine data 10](#_Toc96583546)

# Introduction

When looking at large numbers of paired data, it may be useful to understand if there are relationships in the data. The case under consideration here is protein geometry and questions asked could be: are the bond lengths between the atoms meaningfully related? Or does the backbone angle tau (between the backbone atoms N, CA and C) have any relationship to the backbone dihedral psi (between N, CA, C and N+1)? When we look through these geometric measures manually on a case-by-case basis we sometimes get lucky and find apparently meaningful relationships, or we can make an educated guess as to where to find them, as in the two cases for *Figure 1* where we can imagine a twist around either the N-CA or CA-C bonds and the distance between the atoms on either end correlating with that twist.

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | | a) | b) |   Figure 1 highly correlated geometric plots. They show the relationships between the backbone dihedral angles and the distances between the atoms on either end a) shows phi with C-1:C and b) shows psi with N:N+1. While there is clearly a correlation, how do we know if it is meaningful? |

There is clearly a strong correlation in these plots, and it is clearly meaningful because it represents the relationship that we could predict as the psi and phi dihedrals rotate. It isn’t entirely trivial, because the plots effectively corroborate the belief that the forces in the atoms hold to such a degree that they behave in this geometric way – nevertheless it doesn’t tell us anything we don’t already know. But could it? We believe this is what we expect to see, but that is only an approximate expectation - perhaps this is not exactly the model we already know of, subtle differences could be interesting. Consideration of these plots leads us to 2 questions:

1. Can we find plots like this in an automated way, perhaps some we don’t expect?
2. Can we compare them to idealised geometry to find interesting differences?

The popular statistical methods for correlations are based on linear relationships and we see immediately that in rotating geometry of this kind, linear relationships are not what we need. We therefore propose a new method for identifying correlations in paired data that is distribution free.

The new method is called Williams Divergence from Trivial, and the method looks at the divergence from observed correlated pairs with respect to the unrelated case were they not correlated. The other side of trivial, where the correlation is what we already expect and adds nothing new to our understanding, is calculated by comparison with a synthetic geometric dataset. The metric, known as the divergence, is a calculated as a comparison between the 2d histograms of the two distributions: observed and trivial.

# 1 Williams Divergence from Trivial

The basis of the metric is that if 2 sets of data are correlated with meaning then they diverge from the trivial example where they are a convolution of the 1d histograms.

The most extreme case is of a line y=x, were the x and y values not to be correlated but randomly associated the convolution of them in 2d space would be an even spread.

## 1.1 Example for no variance

An example is described where I have 2500 samples for y=x {x=0,2499}

The data is binned into 5 histogram buckets in 1dimension and convolved into a 5x5 2d histogram. The original scatter plot is binned into a 5x5 2d histogram and the histograms are compared. Figure 2 shows the plots for this example.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | a) | b) | c) | | d) | e) | f) |   Figure 2 Plots for linear example a) is the original scatter plot b) and c) show the 1d histograms for the 2 datasets, d) shows these 1d histograms convolved, e) shows the original data binned into a 2d histogram, f) shows the difference between the convolved and original data. |

The divergence statistic is calculated as follows:

Each 1d histogram bin for both datasets contains a 500 count. As this is convolved it will be even across the 2d matrix with 250,000 in each. However, this is normalised to a total of 1 in the 2d histogram, so each bin contains 250,000 / 6,250,000 = 1/25 or 0.04.

The original data is evenly spread across the 5 buckets that make up the diagonal, so 500 in each bucket, but normalised so each diagonal contains 0.2 and the other buckets contain 0.

Our metric measures the absolute value of the difference of these, and as the upper bound on that is 2, divides by 2 so the metric is between 0 and 1. In this case:

5\* (0.2-0.04) + 20\*0.04 = (0.8 + 0.8) / 2 = 0.8

You can quickly see that the metric is impacted by the bins, and Table 1 shows a few of the calculations with changing bins and sample sizes.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Sample size | 1d bins | Divergence Metric | | 2500 | 5 | 0.8 | | 1000 | 10 | 10\*(0.1-0.01) + 90\*0.01 = 0.9 | | 1000 | 2 | 2\*(0.5-0.25) + 2\*0.25 = 0.5 | | 1000 | 1 | 0, trivially | | 10000 | 100 | 100\*(0.01-0.0001) + 9900\*0.0001=0.99 |   Table 1 Varying parameters for the divergence – the bin size and count size vary to show how the metric can change at the extreme end of no variance. |

Evidently the bin size will impact the value of the metric, so that needs to be taken into account when interpreting it. For data in the extreme case above where there is no randomness at all, the metric converges to extreme correlation as per the formula in *Figure 3*.

|  |
| --- |
| Divergence metric at extreme = 1-  where b = bins on the 1d axis  Equation 1 Formula for the divergence with no variance |

This means that the metric as calculated by this method does not span {0,1) – we will look to normalising the final metric after considering variance.

## 1.2 Trivial examples

Before introducing variance, it is worth looking at examples of trivial relationships. These can occur in 2 ways – randomness of the values means no pattern, or the relationship is not dependent. Figure 3 shows 3 examples where there is no meaningful relationship between the paired data.

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | a) | b) | c) |   Figure 3 Trivial relationships a) and b) have a divergence value=0 and c) has a value close to 0. |

## 1.3 Introducing variance

When there is variance, the bucketing of the observed data is less neat, depending on the bins the observations may be spread across multiple buckets. In this case, the number of observations will impact on the final metric as more observations make smoother bucketing. Figure 4 shows increasing sample size for a line with a mean of x=y in the range{0,20} with a standard deviation of 2. Five bins are chosen to clearly show how the data is distributed.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | a) |  |  | | b) |  |  | | c) |  |  |   Figure 4 Increasing sample size smooths bins, where row a) has 200 samples, b) has 500 samples and c) has 2500 samples. The bucketing becomes smoother so that the higher count converges to the true metric. |

It is certainly not unexpected that the more observations we have the more accurately our calculations will converge to a true metric. It may mean that there is a count below which it becomes indeterminate. If we plot the divergence metric against different sample sizes for 5 bins, and look at different levels of variance, we find the plots in Figure 5: the first which shows how the sample size of 200 is clearly divergent from the others, and the second which shows that the divergence metric converges as the counts increase, with the blue line of no divergence quickly reaching 0.8 (as we expect from our calculations above).

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | | a) | b) |   Figure 5 Sample size improves divergence metric, where a) shows lines for different sample sizes and how as the randomness of the distribution increases they tend to 0. B) Shows that for the different levels of randomess the sample size improves the convergence to the statstic. |

## 1.4 Normalisation of the divergence

The above suggests a stability of the calculated metric, so we can now consider a normalisation based on the bins, knowing that increasing sample size converges to a straight line. The normalisation would be the calculated absolute value metric as per above, divided by 1 – 1/b. This is expressed more clearly in Equation 2.

|  |
| --- |
| Where  i = bin in the x direction  j = bin in the y direction  b = total number of bins in 1 direction  C is the value in the convolved 2d histogram  O is the value in the observed 2d histogram  Equation 2 The Williams Divergence from Trivial metric |

This method is devised to compare paired data with the unrelated case, but it can equally be used to compare any 2 distributions. The trivial, as per the name of the metric, is in the hypothesis of the test. While the primary method of calculating the distribution wrt trivial is to compare it to the unrelated case, any other distributions that you consider to be trivial can be used.

For protein geometry, we are looking at 2 sides of trivial: data with meaningful correlation against unrelated; and data with meaningful correlation against idealised protein geometry.

The python library LeucipPy contains the functionality to calculate the divergence from unrelated trivial as a generic method for any kind of data, free from any need to be protein-geometry related. The idealised protein geometry is specific to this case and although also contained in the library, it is just one possible hypothesis that could be used to look at the other side of trivial.

## 1.5 Divergence p-values

The metric also needs a p-value so we know how unlikely it is that the relationship could have come about by random chance. For this, being distribution free, we need to perform a relatively computationally intensive re-sampling for every single pairing that we consider.

The p-value is calculated by resampling with replacement each of the datasets independently, and then randomly shuffling that data so that they are no longer paired. The p-value is then calculated to be the upper tail, where the standard hypothesis would be that a p-value of <0.05 would be significant. *Figure 6* shows the scatter plots and histograms of the calculated metric for 3 distributions of increasing variance. The combination of p-value and divergence give a clear indication of correlation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  |   Figure 6 The p-value calculated for 3 distributions where row a) is an extreme correlation, b) has a line with some variance, and c) is completely random. They were calculated with 500 re-samplings. |

The p-value calculation is only available for the comparison to trivially unrelated as part of the library calculations, not to the other side of trivial – when a supplied alternative distribution is given only the metric is calculated.

## 1.6 Results

The usefulness of this metric is perhaps best evaluated by the results, and *Figure 7* shows the top 16 correlations found for a set of high resolution protein structure data (<= 1Å) for glycine only residues. The bottom row shows the 4 least correlated pairs, and you can see the p-value shows they are not significant.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |   Figure 7 Top diveregnces calculated for high resolution glycine data, the figures are small but clearly demonstrate that the metric has picked out those highly correlated. For the top 16, the p-values are all zero, the correlations are highly significant. For the bottom 4 the p-values are not significant or significantly even less correlated. |

(not done yet)

(And, in the above there is a bit missing about when the statistic is important, though we have p-values. But when it is normal about a line the different bins have different values.)

# 2. Idealised Protein Geometry

## 2.1 Creating a synthetic model

Creating a model of an idealised set of residues where turns are trivial based on expected random bond length and angle variation.

## 2.2 Idealised plots Comparison

Guven the idealised model, compare it to the plots and finds any that are not trivial.

And thus we have something interesting.

## 2.3 Results

Results showing both the geometry and the metric and p-value and how it sorts data

## Discussion

And conclusions on all the above!

## Future Direction

Resolution, impact of refinement, use of the correlations for validation, psi and phi and tau.

# References

<https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained>

<https://www.uio.no/studier/emner/matnat/math/STK4400/v05/undervisningsmateriale/A%20distribution-free%20approach%20to%20rank%20correlation.pdf>

<https://pages.stat.wisc.edu/~mchung/papers/huang.2019.ISBI.pdf>

<https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence>