

Applied Statistics

A statistical approach in protein stability prediction

Here I explore a method for rapid prediction of protein stability upon change of residue composition. A bayesian probabilistic knowledge-based state of the art multibody multinomial model, MuMu, was investigated as a mean of predicting protein structure stability of mutated proteins in real-time. By using a readily available method for predictions of protein side-chains it was possible to develop a high-throughput method for mutating proteins and determining their stability by MuMu. The probabilities given by the MuMu model was compared to a dataset of 458 experimental $\Delta\Delta G$ values originating from 14 proteins. I seek to increase the MuMu models correlation to experimental determined $\Delta\Delta G$ values by investigating the data with statistical analysis. The dataset was divided into a series of subsets based on the mutated amino acids physical properties. None of the methods, however, did not reveal a significant relationship between the probabilities and $\Delta\Delta G$ values.

Introduction

Previous analysis of the MuMu model has shown that there is a correlation between MuMu probabilities and $\Delta\Delta G$ values. Previous analysis was done on a single protein with 95 mutations. See figure 1.

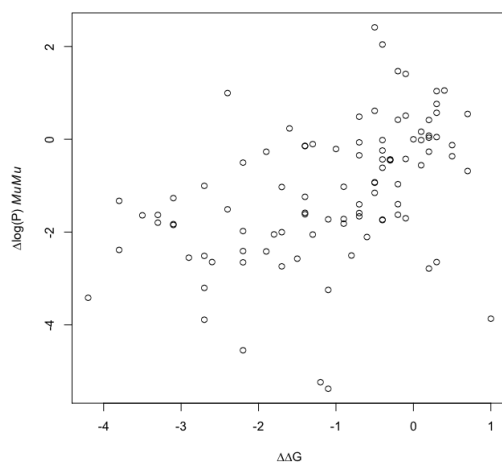


Figure 1. Data is not perfectly centered around a diagonal line. y-axis shows the change in logarithmic probability upon mutation, $\Delta\log(P(A|X))$. x-axis shows the change in change of free energy upon mutation.

A Pearson correlation analysis revealed a correlation of $r = 0.45$ and a hypothesis test

rejected a null hypothesis of no correlation. $r = 0.45$ indicates a modest correlation (Taylor, R. 1990). However, the correlation is relatively high in terms of protein stability prediction. An increase in data size is naturally the next step to further assess the reliability of the MuMu model. The probability data was made by introducing mutations to known pdb structures of the proteins from where the $\Delta\Delta G$ values were obtained. The mutations were done by a program called SCWRL4. Next the probabilities of the wild type protein and all the mutated versions of the protein were determined by MuMu. All this was put together by a python script - allowing fast introduction of mutations and their influence on the probability. This influence was represented by a change in probability from wild type protein to mutated protein. A dataset of 95 $\Delta\log(p)$ values.

In this project I rewrote the python script to allow for copy-pasting protein stability data, in $\Delta\Delta G$, directly from:

“<http://www.abren.net/protherm/>“,

which is a thermodynamic database. From that I got 458 $\Delta\Delta G$ values. I generated 458 $\Delta\log(p)$ values from the structure of the corresponding proteins. All this was stored as a pandas data frame so the data could be easily accessed and manipulated.

Analyzing the data

First I wanted to do a simple scatter plot as previously done. I wanted to make sure that my data was made correctly so I tried to replicate the data seen in figure 1. I ran the python script with the same protein, namely 1STN. The same plot was generated, not shown here. I then proceeded to generate the complete dataset from 16 proteins. The data was filtered of Null values and I ended up with 458 values of $\Delta\Delta G$ and 458 $\Delta\log(p)$ values. I manually assessed the data to make sure that there were no errors.

I then proceeded to make the scatter plot, figure 2.

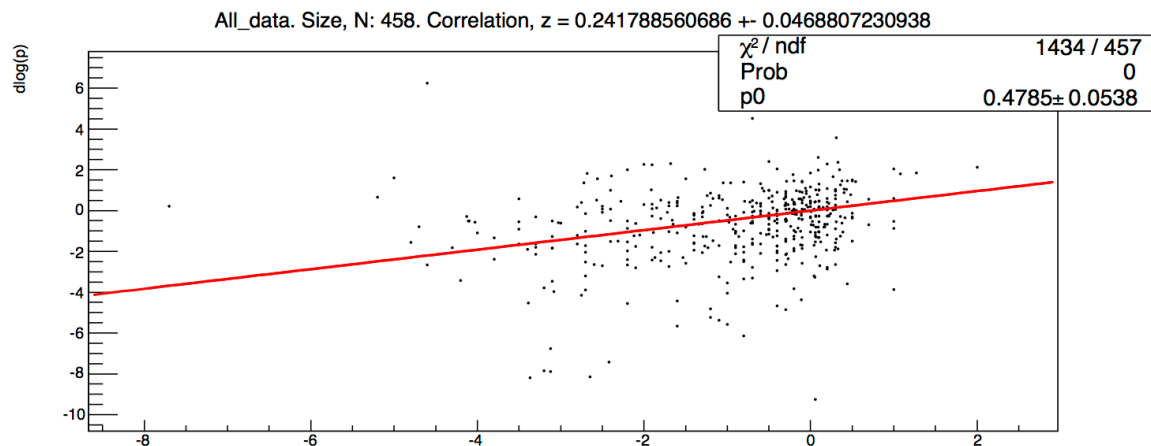


Figure 2. Scatter plot of $\Delta\Delta G$ values, x-axis, and $\Delta\log(p)$ values, y-axis (complete data set). Data is fitted with a linear function.

The plot was fitted with a linear function. The assumption being that no change in probability would mean no change in stability and thus no change in $\Delta\Delta G$. This means that the fit was without extra parameters. The first impression by looking

at the plot is that there may be some, however little, correlation in between $\Delta\Delta G$ and $\Delta\log(p)$. The $\Delta\log(p)$ was calculated so that a negative value would indicate increasing stability which is the same for $\Delta\Delta G$ values. The chi-square of the fit is 1434 and ndf is 457. This does not indicate a good fit and when we look at the probability that is extremely low we may conclude that it is indeed a bad fit.

I wanted to test a null hypothesis that there is no fit against an alternative hypothesis that there is a good fit. To do this I made a runs test. This revealed an extremely high sigma of 10.68 and the run test gave a value of 4.48 sigma. This means that we are more than 4 sigma away. The alternative hypothesis can therefore be rejected.

The correlation, rho, was calculated using ROOT. The correlation coefficient can be transformed into the variable z, as the sample size is not *very large*:

$$z = \frac{1}{2} \ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right).$$

z has a more gaussian shape than rho and the standard derivation of z is:

$$std = \frac{1}{\sqrt{N-3}}$$

This was calculated by ROOT and attached to all figures provided, see figure 2. The correlation coefficient of 0.25 is very low.

Different amino acids have different properties. It may be that the MuMu model is better at predicting changes in stability when the mutation is of a certain type. The dataset was split up into different categories to investigate this. First I choose to look at the most common classification of amino acids. I generated a python script that could

+/- 0.11 (table 1 & appendix figure 1). It would seem that they have a high contribution to the correlation of the data. A manual inspection of the data reveals that the data does seem somewhat random it is not clear if that is the case without collecting more data. The correlation coefficient for changes of polar amino acids to hydrophobic amino acids is 0.39 +/- 0.12. It does look like there is a linear tendency. MuMu may be better at predicting stability when there are mutations from hydrophilic

	z	error
All data	0,24	0,05
Mutated to charged amino acid (figure app. 1B)	0,44	0,11
Mutated to uncharged amino acid (figure app. 1C)	0,20	0,05
Mutated to aromatic amino acid (figure app. 1D)	-0,2	0,24
Uncharged amino acid mutated to charged amino acid (figure app. 1E)	0,25	0,08
Charged amino acid mutated to uncharged amino acid (figure app. 1F)	0,65	0,15
Mutated to hydrophobic amino acid (figure app. 1G)	0,22	0,06
Mutated to hydrophilic amino acid (figure app. 1H)	0,15	0,1
Hydrophobic amino acid mutated to hydrophilic amino acid (figure app. 1I)	0,43	0,17
Hydrophilic amino acid mutated to hydrophobic amino acid (figure app. 1J)	0,4	0,12

Table 1. Correlation coefficients, z and their errors. Based on subsets of the complete dataset from appendix, figure 1.

filter the data based on different criteria. Figures are shown in are found in appendix, figure 1. Correlations from the fits are shown in table 1.

The plot with charged amino acid to uncharged (peptide) amino acid and the plot with the charged amino acid have relatively high correlation (speaking in terms of structure prediction), 0.65 +/- 0.14 and 0.44

to hydrophobic. It seems that mutations that lead to aromatic amino acids has a negative overall influence on the correlation. However it is inconclusive if this is the case due to the low sample size of 19.

The MuMu model is trained on non-membrane proteins. All proteins chosen here are therefore non-membrane. It may be that MuMu is more sensitive to different

kind of proteins and that it is better at predicting the stability of some of the chosen proteins than others. This time I split up the data into subcategories that are the proteins. Same analysis of correlation was done. See appendix, protein figures. The correlation coefficients can be seen in table 2. There is negative influence on the correlation coefficient from a few of the protein, especially 1VQB, 1CYO. 3SSI have a perfect correlation however the plot shows that it is caused by a very bad fit. 1STN has about the same correlation as seen in the previous study. 1PGA and 1BVC has a high correlation but the standard derivation is also quite high thus making it unreliable. 1ROP has the highest correlation even with

Conclusion

A recurring thing is the need for a larger sample size. This is a problem especially when looking at the subsets. There is an indication that MuMu might be better at predicting stability of proteins with mutations of hydrophobic amino acids to hydrophilic. The fit for the entire dataset is not good. The predictions are not consistent enough to use it as a way of determine $\Delta\Delta G$ values on the basis of MuMu values. That is with the current dataset.

Literature

	z	error
All data	0,24	0,05
1BNI	0,23	0,08
1BPI	0,03	0,25
1BVC	0,63	0,41
1C90	0,28	0,29
1CSP	0,17	0,25
1CYO	-0,32	0,41
1FTG	-0,1	0,41
1IGV	-	-
1PGA	0,64	0,41
1ROP	0,75	0,25
1STN	0,47	0,11
1VQB	-0,2	0,25
2LZM	0,29	0,12
2RN2	0,16	0,32
3SSI	1	0,45

its error of 0.25. The sample size of 1ROP is not large enough to say anything conclusive.

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, 6(1), 35–39. doi: 10.1177/875647939000600106

Roger J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*.

Appendix

