The University of Akron
IdeaExchange@UAkron

Honors Research Projects

The Dr. Gary B. and Pamela S. Williams Honors College

Spring 2015

Evaluation of the Signature Molecular Descriptor with BLOSUM62 and an All-Atom Description for Use in Sequence Alignment of Proteins

Lindsay M. Aichinger
The University Of Akron, lma36@zips.uakron.edu

Evaluation of the Signature Molecular Descriptor with BLOSUM62 and an All-Atom Description for Use in Sequence Alignment of Proteins

Lindsay Aichinger

Department of Chemical & Biomolecular Engineering

Honors Research Project

Submitted to

The Honors College

Approved:

Date 5/1/15
Honors Project Sponsor (signed)

Donald P. Visco, Jr.
Honors Project Sponsor (printed)

Date 5/1/15
Reader (signed)

Nic D Leipzig
Reader (printed)

Jie zheng Date 5/4/15
Reader (signed)

JIE ZHENG
Reader (printed)

Accepted:

Date 4 MAY 2015
Department Head (signed)

H. MICHAEL CHEUNG
Department Head (printed)

Date 4 MAY 2015
Honors Faculty Advisor (signed)

H. MICHAEL CHEUNG
Honors Faculty Advisor (printed)

_________________ Date _________
Dean, Honors College

EXECUTIVE SUMMARY

This Honors Project focuses on furthering the study of using predictive mathematical models to determine if one amino acid can be substituted for another. Substitution of amino acids allows drugs to be developed using amino acids as substitutes for the targeted amino acid. A method to predict the success of this substitution mathematically would allow drugs to be developed faster and have a greater success rate when then used in human trials. This Honors Project focused on three aspects of this topic:

- Duplicating the work of Jean-Loup Faulon who used the molecular signature to investigate drug effectiveness
- Comparing the molecular signature kernels to three of the BLOSUM matrices (30, 62, and 90) to test the accuracy of the mathematical model
- Manipulating the kernel matrix in order to improve the relationship:
    - 1. Focusing on side groups
    - 2. Changing how the structure was represented in the matrix by increasing the initial height distance from the central atom (Height 1 and Height 2 included)

The molecular signature is a method that describes a chemical or biological structure by counting individual structure groups based on distances from a certain atom. The quantifiable approach is used to describe chemicals and will now be used to describe amino acids and proteins. If the approach is proven to be reliable, the molecular signature could be used to predict whether a drug will be effective inside a patient due to similarities found between the drug and the target protein. There were multiple design constraints for this project. The first was the comparison with the BLOSUM matrices (30,62, and 90) which characterized the substitution ability of each pair of amino acids as found in nature. The next constraint was the use of the molecular signature as a way to describe the structure of the amino acids. The last constraint was the usage of the signature kernel equation which was used in order to repeat a previous study by Jean-Loup Faulon.

The Honors Project first starts by calculating the kernel values with the molecular signature values for the original matrix (without any modification) and resulted in a linear correlation R^2 value with BLOSUM62 of 0.83. In order to improve the linear relationship, it was predicted that the side groups for the amino acids were the most important aspect of the structure. In order to investigate that, the matrix was modified to take out the signature values that described the “backbone” of the amino acid and focused on the side groups. The linear relationship of the BLOSUM62 and the kernel values of the modified matrix did improve with a higher linear correlation R^2 value of 0.87. Other results included trends in regards to the amino acid and the BLOSUM matrix values. According to this investigation, the amino acids with large nonpolar or large aromatic side groups resulted in the lowest BLOSUM scores. For example, Tryptophan, Arginine, and Phenylalanine were among the amino acids with the lowest BLOSUM scores..

Further investigation with regards to atom height, was attempted in order to improve the linear relationship more. The next experiment tested a matrix with a height of 2 with molecular signature. The linear relationship of the BLOSUM62 and the kernel values of the height 2 matrix did improve even more and resulted as 0.93. According to this investigation, the improved relationship will occur when there is a larger height with an all-atom approach.

As a result, it can be concluded that improvement of the molecular signature kernel matrix should not rely on one aspect of the amino acid. The prediction of similarity as seen in nature relies on many complicated factors but polarity should not be discounted as one of them. From observations of the results, it is apparent that polarity does make a difference in nature and using height 1 or a two-dimensional method will limit the results. Other important properties could be bulkiness of the side group, three-dimensional structure, and the orientation of amino acid. As the height of the molecular signature increases, the three-dimensional structure and orientation of the amino acid will be represented as shown by the higher linear relationship.

There are many important implications of this investigation. The following is a list of skills obtained from the project: increased knowledge of the bioinformatics field and knowledge of ways to evaluate mathematical similarity methods, a better understanding of structures of amino acids and how the properties of proteins directly correlate to those building blocks, and finally more proficient programming skills and understanding of the direct application of C++ in order to solve a big data problem. The project improved self confidence in the subject matter and it gave the opportunity to work on a project with the help of a senior advisor. Molecular signature descriptors could prove to be a reliable way to quantify the relationship between amino acids within proteins. If this became a reliable method it could be an affordable way to predict protein –protein interaction and give an early start to investigate drug options for patients. Safer drug use will result in less people dying every year from drug use complications. Also, a future benefit could be the identification of dangerous drug to drug interactions thus helping save lives from unknown negative drug interactions.

Moving forward, future investigations should account for height 3 or higher and investigate different trends that mimic the trends found within the BLOSUM62 matrix. If further patterns are found within the BLOSUM62 matrix then those could be used to modify future matrices and find an even higher linear correlation than was found in this investigation. It is also recommended to keep in mind that multiple factors (different pH levels and intermolecular forces) occur at once and individual investigations may overlook the synergistic effect on the structure of the amino acid as seen in nature.

BACKGROUND:

The research and investigation of this Honor's project primarily revolved around a paper from Jean-Loup Faulon et al called *Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor*[1]. Faulon's research focuses on possible bioinformatics application of the molecular signature. The molecular signature is used to describe chemicals but further application includes describing amino acids and proteins. Faulon's investigation tries to use the molecule signature to predict similarities between a drug and its target protein. The implication for the future is to use this method to predict if a drug would be effective for the patient prior to the patient ingesting the drug. In a section of Faulon's research, he investigates the relationship between pairs of amino acids calculated by the molecular signature and the values of the substitution BLOSUM62 matrix. This relationship is important since it is used to evaluate the effectiveness of the molecular signature approach.

The BLOSUM 62 matrix numerical values are determined by two strands of DNA that are 62% similar in sequence alignment. The matrix values represent how likely two amino acids can be substituted for each other. If the BLOSUM62 matrix element for a certain pair has a large positive value then it occurs more frequently in nature and is less likely to impact protein function. For example, Valine and Isoleucine pairing have a BLOSUM62 score of 3. Proline and Proline have a score of 7 since their structure is the same. Some pairs of amino acids can have values as low as -4 or -3 which means these substitutions would rarely occur in nature and if they did occur would likely result in an important change in protein function. For example, Valine and Tryptophan have a score of -4, which is due to a difference in structure since Valine has a nonpolar side group while Tryptophan has an aromatic side group.

The molecular signature is an all atom approach used to quantify the relationship between two chemical structures or, in this case, amino acids. The atomic signature is a basic method for identifying an amino acid. Due to the growing databases being used for the identification of different proteins and chemicals, there is a need to find a highly reliable computational method for categorizing chemical structures. Non-computational methods include high throughput screening and virtual screening. High throughput screening uses known groups of compounds and compares their structure to the unknown structure in order to identify the unknown. Virtual Screening utilizes homology similarities in order to identify the unknown compound. When identifying proteins with virtual screening, ligand structure is highly important.[2] The computational method is a better approach since it is faster and more accurate and the other approaches involves experimentation.

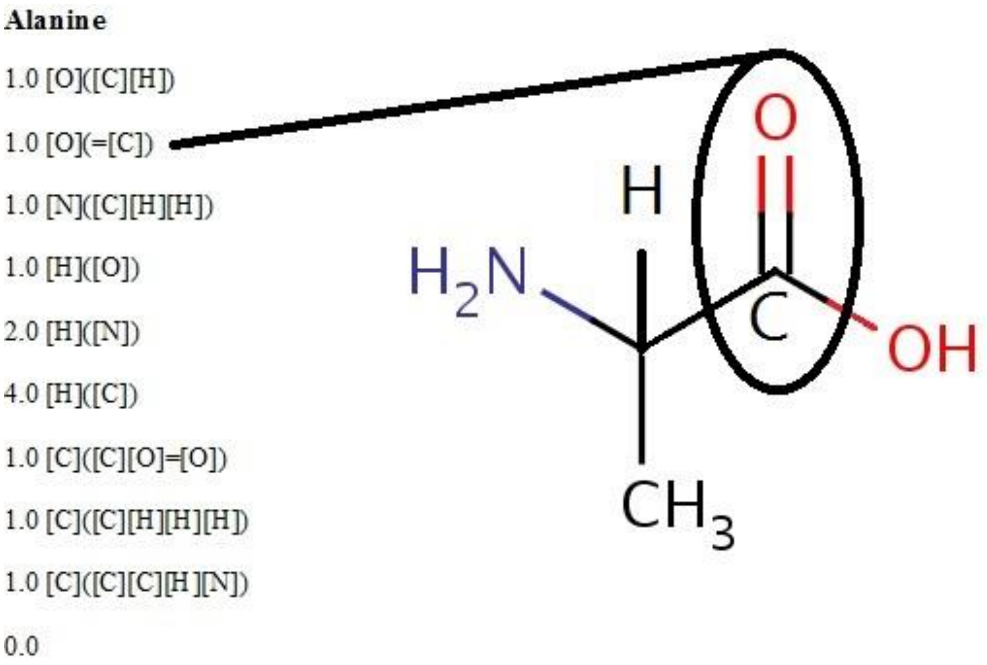
There are also other approaches that build upon the molecular signature in order to describe properties. One example is QSPR that is used to describe the properties of different polymers.[3] In order to expand upon the molecular signature approach; it needs to be proven reliable at its basic level, meaning usage at Height 1 or Height 2. The molecular signature descriptor is a method that is public and free to use around the world. There is a strong motivation to improve this approach in order to expand upon the predictability of molecular structures. The implication is to use this approach to predict the usefulness of a drug by comparing the drug protein and the target protein expressed by the patient’s genome. The individualized approach could advance personalized medicine and potentially save millions of lives.

In Faulon’s research, he investigates the molecular signature and how it is used to quantify the similarity between two amino acids. The molecular signature of a molecule is comprised of atomic signatures that are denoted by the amount of atoms which are a predefined distance from a central atom. The predefined distance is called the signature height that is characterized by a vector. In this Honor’s project first investigation, the height distance was 1 which was later increased to a height of 2. In Faulon’s research, he used a height of 1 as well. If G is a function of V (vertex atom) and E (signature height), then Equation 1 shows the equation for the molecular signature:

$$^{h}\sigma(G) = \sum_{x \in V} {}^{h}\sigma_G(x) \qquad (1)$$

Equation 1. $^{h}\sigma_G(x)$ is the unit base vector that is the atomic signature G embedded at x with a height h.[1]

For example, Figure 1 shows the molecular signature for the matrix row of Alanine. The structure of the amino acid is shown next to it. The signatures are generated by focusing on atoms right next to each other. This process is utilizing a height of 1. The different atom combinations are counted and used in the matrix row.



Alanine

1.0 [O]([C][H])

1.0 [O](=[C])

1.0 [N]([C][H][H])

1.0 [H]([O])

2.0 [H]([N])

4.0 [H]([C])

1.0 [C]([C][O]=[O])

1.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])

0.0

Figure 1. The molecular signature matrix row for Alanine.[4]

While signatures are vectors, Faulon's research also included calculations of signature kernels where the results are scalar values. The signature kernel represents a relationship between two amino acids with the values ranging from 0 to 1. A kernel value that is closer to one means that the structures are more closely related. The kernel is calculated by taking the dot product of the molecular signatures in the numerator. The denominator is for normalization and is found by squaring each value in the vector and summing them, and then taking the square root. An example of a calculation of the kernel is found in Appendix D. Equation 2 shows the equation for calculating the signature kernel.

$$^{h}k(A,B) = \frac{^{h}\sigma(A)\cdot\, ^{h}\sigma(B)}{|\, ^{h}\sigma(A)||\, ^{h}\sigma(B)|} \qquad (2)$$

Equation 2. Kernel function with the denominator as the vector norm.

where $^{h}\sigma(A)$ and $^{h}\sigma(B)$ are the molecular signatures from Equation 1 and $|\sigma(A)|$ represents the norm of $^{h}\sigma(A)$.

Faulon's research included a figure that showed the correlation between the calculated values for the signature kernel of pairs of amino acids and the BLOSUM62 value.

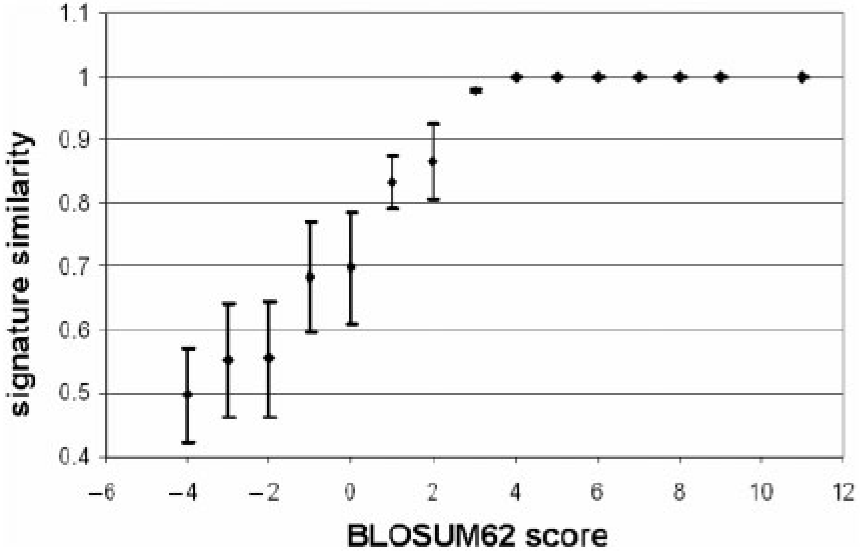


Figure 2. The signature kernel vs the BLOSUM62 score. Standard deviation values were included.[1]

Figure 2 shows a strong correlation between the signature similarity and the BLOSUM62 score.

PURPOSE:

The purpose of this Honors Project is three-fold. To replicate Faulon's research and compare results with more than one BLOSUM matrix. To investigate how concentrating on the structure of the side chain of each amino acid affects the results with the three BLOSUM matrices (30,62, and 90). The signature kernel matrix will be manipulated to only include the structure of the "side chain". Also, the kernel matrix will also be changed to increase the height distance from a central atom to include height 2 . The last two changes will involve manipulation of the signature kernel matrix. It was predicted to improve the linear relationship when compared to the BLOSUM62 matrix. The results were then compared in order to determine if there is a higher correlation focusing on an all-atom approach (Height 1 and Height 2) or focusing on the structure of the side chain with the BLOSUM matrices.

METHODS:

A program was written in C++ in order to calculate the signature kernel. The molecular signature was input as an array with 20 rows (20 amino acids) and 34 columns describing their structure. It was calculated using 2d mol files from the NIST WebBook of each amino acid. The molecular signature for each amino acid can be found in Appendix B while the actual matrix is found in Appendix C. Each row was fed in and multiplied by each other using the kernel equation (Equation 2) from the Background section. An example of that calculation is found in Appendix D. The kernel output was also an array that was analyzed using an Excel spreadsheet. Appendix A shows the code used by the IDE compiler CodeBlocks.

RESULTS:

The signature kernel followed a fairly similar pattern with the BLOSUM matrices. In Figure 3, the BLOSUM matrices 30, 62, and 90 and their respective kernel signatures were compared against each other. BLOSUM62 matrix and the signature kernel had the highest linear correlation value (R^2 value was 83%) followed by BLOSUM90.

Figure 3 shows how there are patterns associated with the BLOSUM matrices and the signature similarity. There is a general positive correlation between the signature value and the BLOSUM score. The R^2 linear correlation value was 83%.

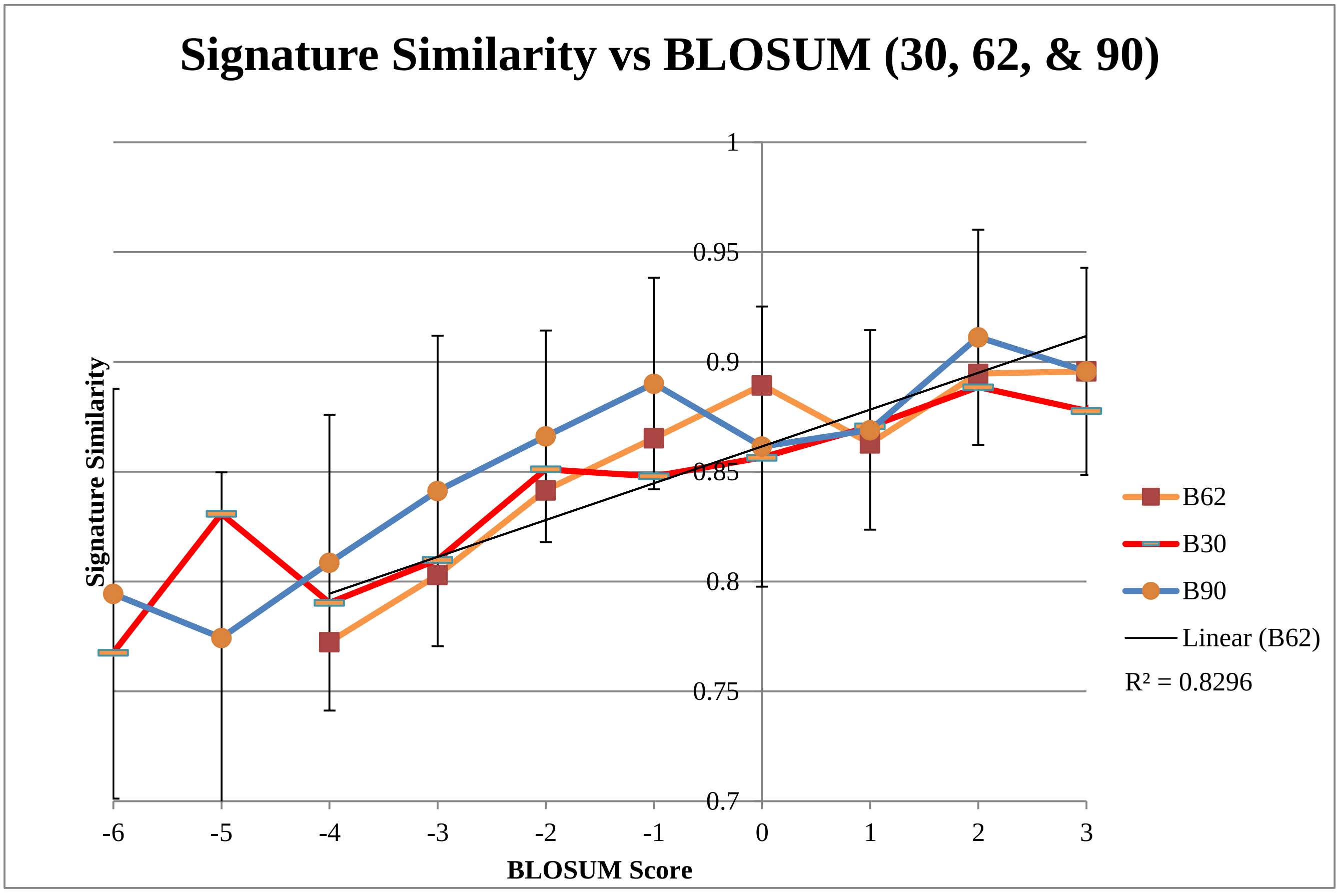


Figure 3. Signature similarity as seen in Faulon's research with standard deviation ranges.

Further exploration investigated the relationship of the BLOSUM score and the individual amino acids. The purpose of the further investigation is to find patterns between BLOSUM62 scores and amino acids. These patterns will help find possible ways of improving the signature kernel matrix. Some amino acids averaged lower scores than others due to their difference in structure. BLOSUM62 appeared to be the most accurate correlating a linear relationship between the BLOSUM score and the signature kernel value and so it was the main focus of the three BLOSUM matrices. Table 1 shows that the individual size of the amino acid influences the associated BLOSUM score. The number of times an amino acid had a certain range of BLOSUM scores were counted to compare the amino acids individually. Each amino acid has a total of 19 values across Table 1 since it can be paired up with all of the others. This table shows how each amino acid varies with the other. For instance, Tryptophan and Phenylalanine are very large amino acids with benzene rings in their structure. These amino acids are largely associated with negative BLOSUM scores. Interestingly, Isoleucine is also associated with a large negative BLOSUM score as well. Isoleucine is an amino acid that is nonpolar that could cause disturbances when substituted for amino acid with different properties. The observations from Table 1 show that structure alone cannot predict the substitution ability of an amino acid and that polarity is also a factor.

Table 1. Individual Amino Acids with associated BLOSUM62 matrix score.

		BLOSUM62 Score			
		-4 and -3	**-2 and -1**	**0 and 1**	**2 and 3**
INDIVIDUAL AMINO ACIDS	**Alanine**	1	13	5	0
	Arginine	5	9	4	1
	Asparagine	6	4	9	0
	Aspartic Acid	8	7	3	1
	Cysteine	9	9	1	0
	Glutamic Acid	3	8	7	1
	Glutamine	5	7	5	2
	Glycine	7	9	3	0
	Histidine	4	10	4	1
	Isoleucine	10	5	2	2
	Leucine	6	9	2	2
	Lysine	4	10	4	1
	Methionine	2	12	4	1
	Phenylalanine	8	6	4	1
	Proline	6	13	0	0
	Serine	1	10	8	0
	Threonine	0	15	4	0
	Tryptophan	10	7	1	1
	Tyrosine	3	13	0	3
	Valine	6	8	4	1

Table 1 shows evidence that the entire structure of the amino acid should not be the only tool to predict the substitution ability. Figure 4 shows a graph where the only inconsistent parts of the kernel signature values were the end points (0.65 and 0.950-0.999). A lack of linear relationship can be seen at both ends of the figure. If the specific polarity is accounted for on the side chains, then the prediction of the signature kernel would be more accurate. In order to help interpret the data in Figure 4 and Table 1, Appendix B shows the structure of each amino acid and the classification group it belongs to based on its side group structure. Appendix B also shows the molecular signature of each of the structure of the amino acid.

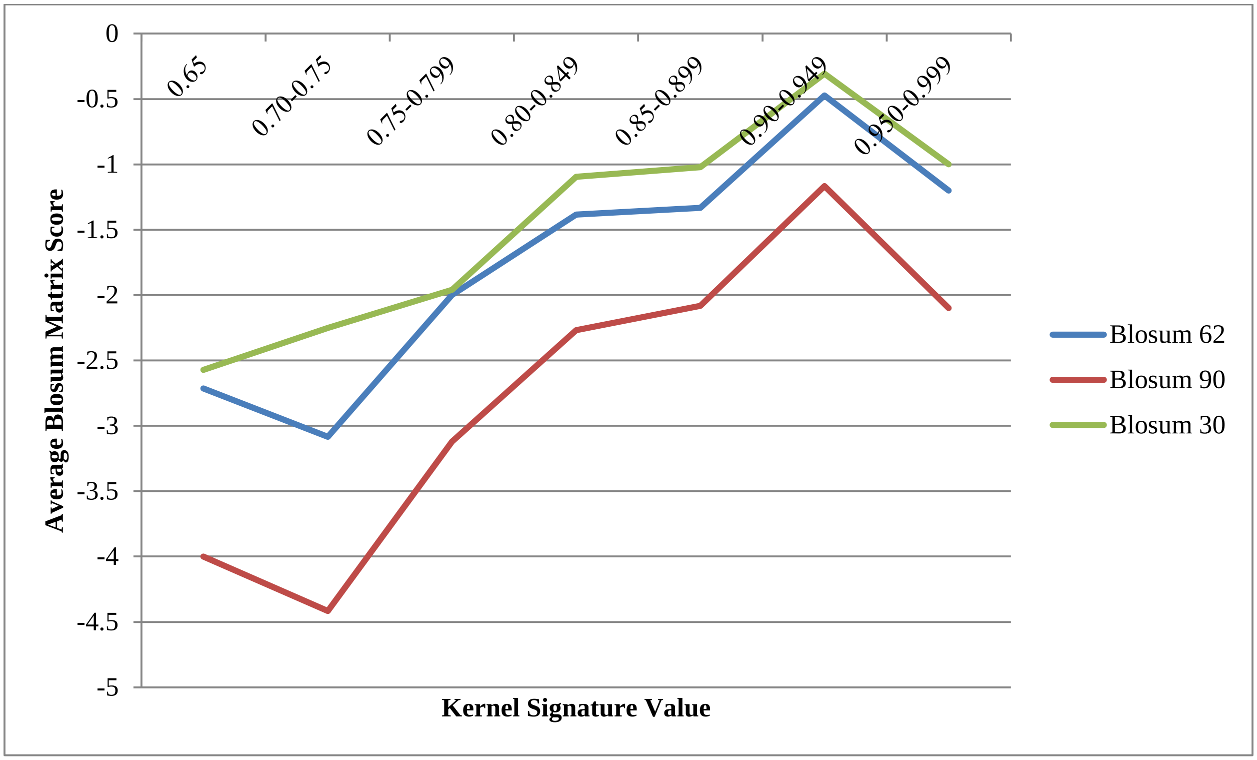


Figure 4. Line graph showing the association between the average BLOSUM scores of the three BLOSUM matrices and the kernel signature values.

FURTHER RESEARCH: Improved Molecular Signature

The signature kernel used in the investigation quantifies the molecule in an all atom approach which means that every part of the atom counts equally to calculate the signature kernel value. If polarity is a factor in terms of substitution ability, the atoms in the side chain should have a higher influence on the signature kernel values. A modified version of the molecular signature matrix was used in the next analysis in order to obtain a more precise relationship between the molecular signature and the BLOSUM matrices. The modified matrix values can be found in Appendix C. Figure 5 shows the new relationship of the BLOSUM matrices and the modified matrix molecular signature kernel. Between Figure 3 and Figure 5, there is an improvement in the R^2 values. The R^2 values increase from 0.83 to 0.87 thus showing a higher correlation.

Additional investigation was done in order to calculate the molecular signature kernel values at height 2. It was hypothesized that this change would result in even better accuracy or a higher linear relationship with the BLOSUM 62 matrix. This is because at height 2, the perimeter or outside atoms are more represented in the kernel values which better mimics the structure of the amino acids in nature. As the height values increase, so does the representation of the three-dimensional space of the amino acid. A higher height difference should improve the linear correlation R^2 value.

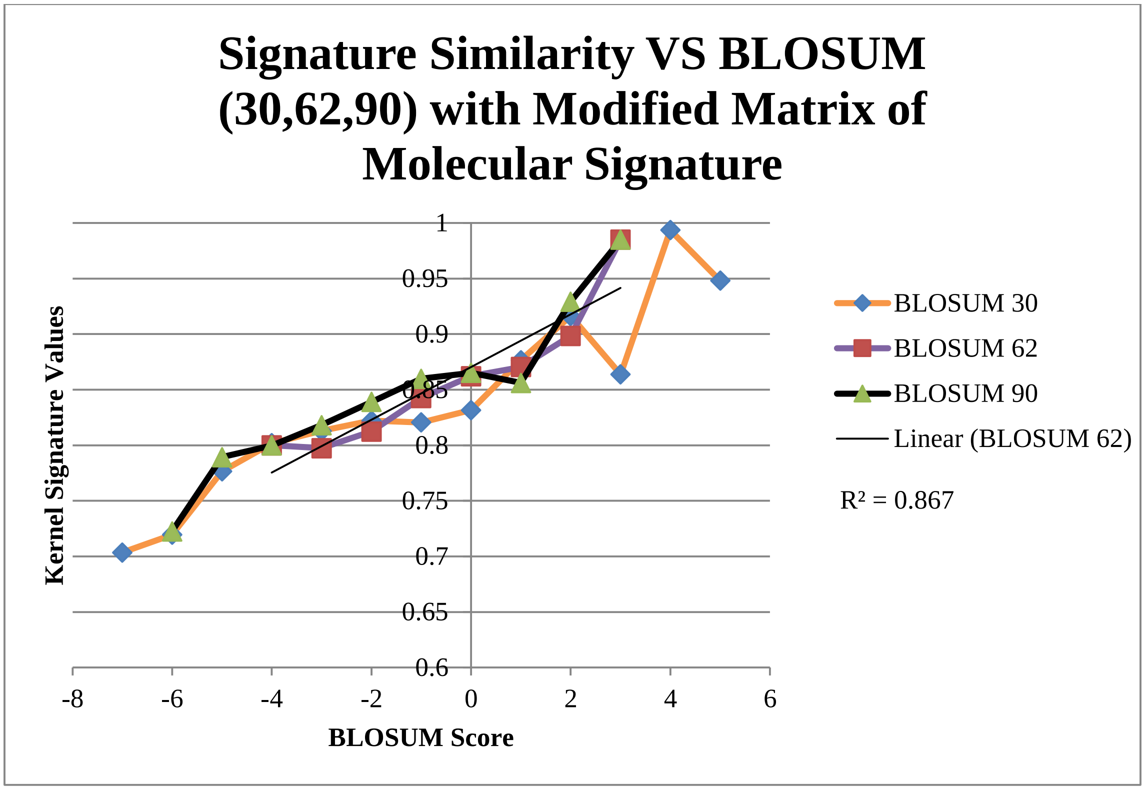


Figure 5. Signature similarity with the modified matrix for Molecular Signature.

Also, in addition, an investigation of molecular signature at height 2 was calculated and compared to the original matrix. Figure 6 shows the results from the investigation. As expected, the linear relationship improved increasing from 0.83 in Figure 3 to 0.93 in Figure 6.

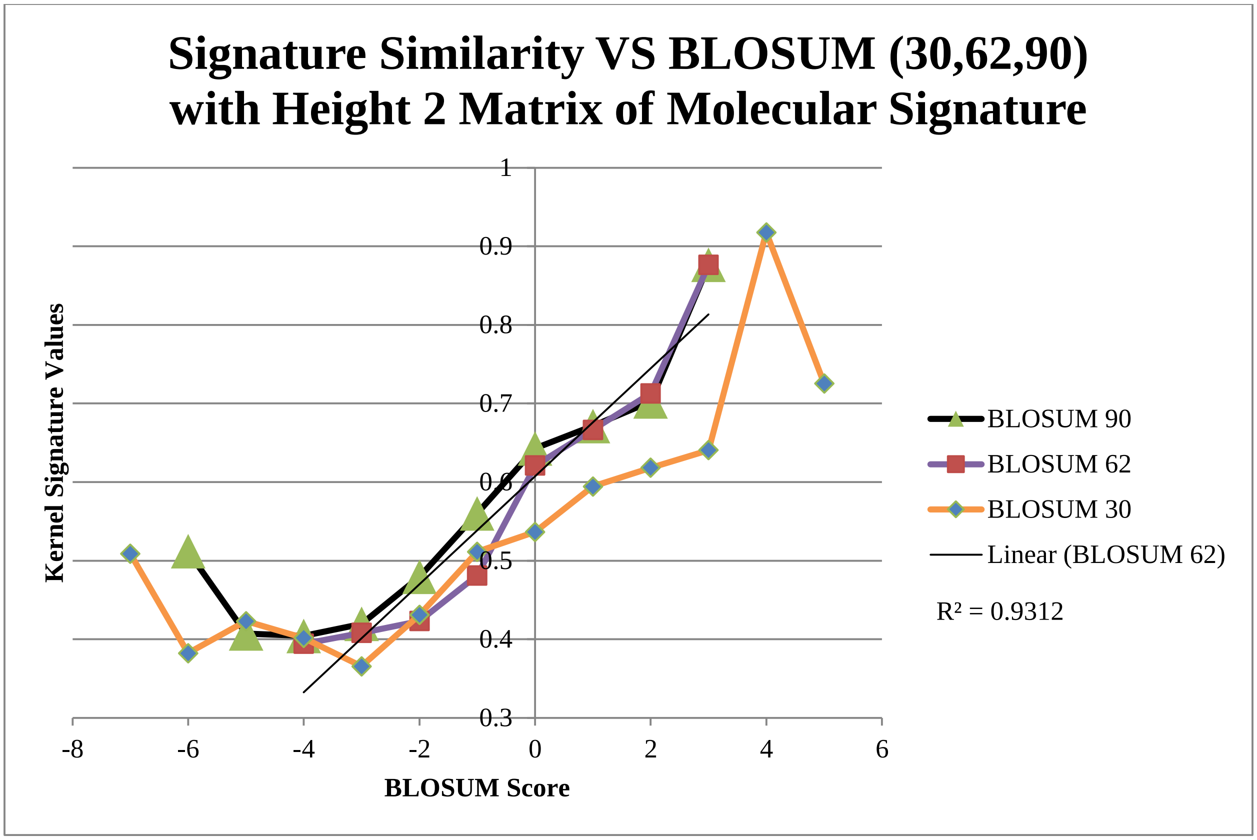


Figure 6. Signature similarity with the Height 2 matrix for Molecular Signature.

Figure 7 shows the final result of the modified signature matrix kernel calculations using the BLOSUM matrices 90, 62, and 30. The trends associated with this figure shows that there is an improvement with the correlations from Figure 4 to Figure 7. In Figure 4, the middle of the graph had a higher correlation than the ends of the graph that showed a slight dip in the data. In Figure 7, the ends of the graph included less “dips” in data while the middle of the graph tended to oscillate more than in Figure 4.

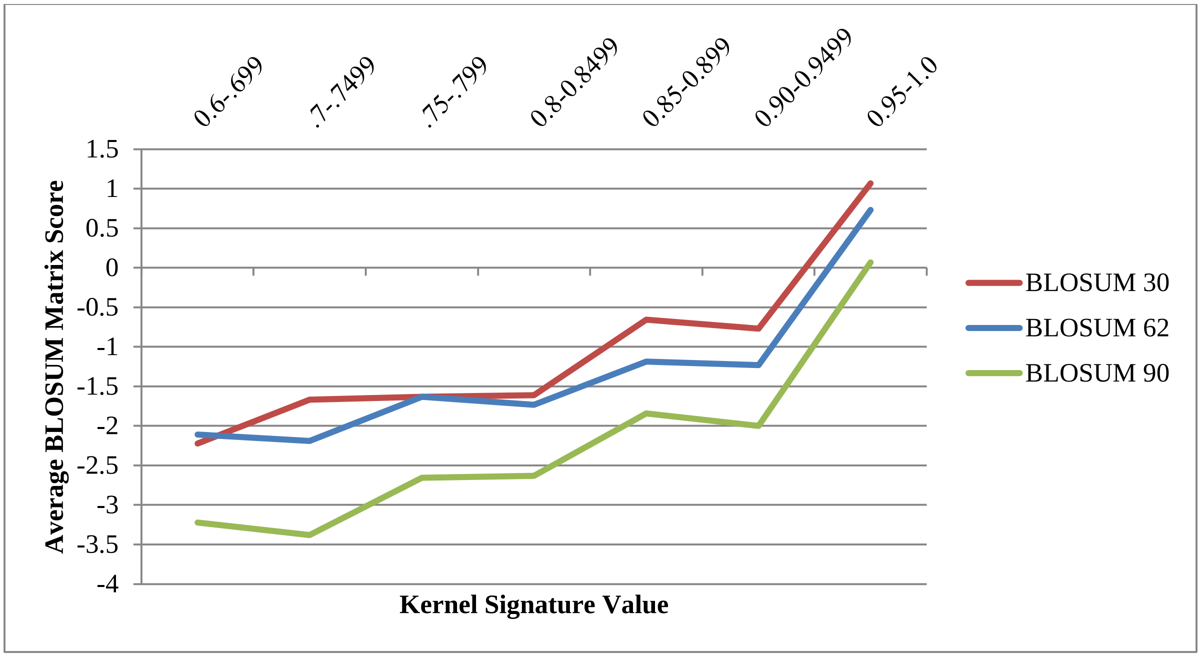


Figure 7. Line graph showing the association between the average BLOSUM scores of the three BLOSUM matrices and the modified matrix kernel signature values.

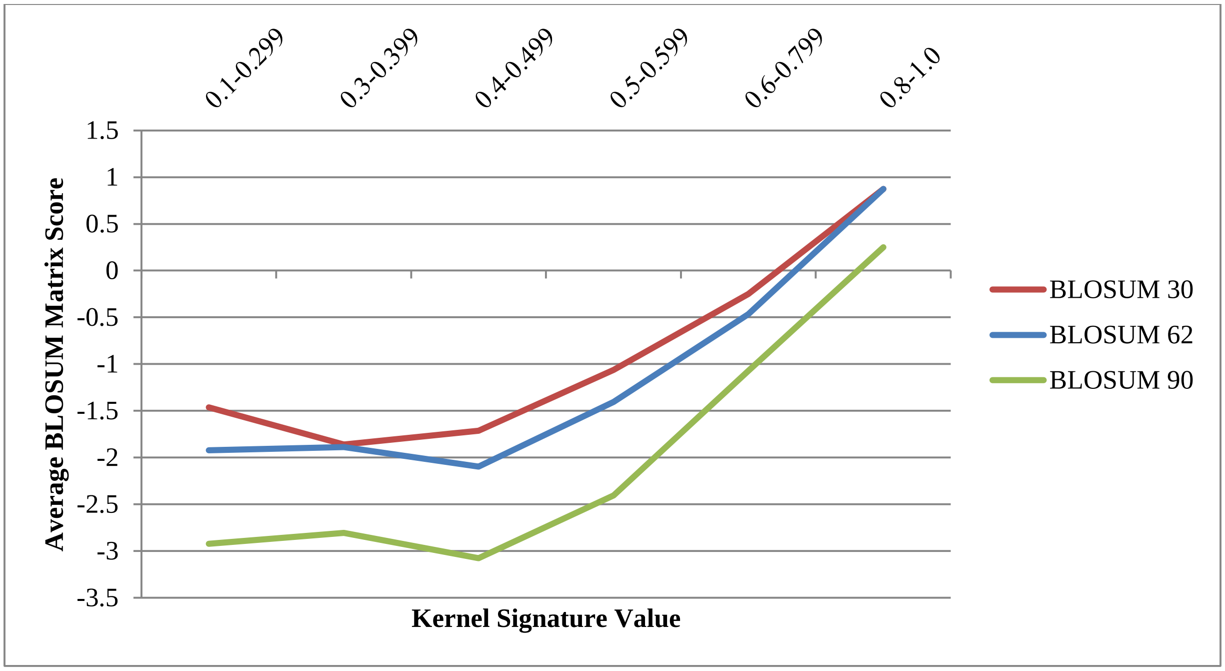


Figure 8. Line graph showing the association between the average BLOSUM scores of the three BLOSUM matrices and the height 2 matrix kernel signature values.

Figure 8 shows the final result of the height 2 signature matrix kernel calculations using the BLOSUM matrices 90, 62, and 30. The trends associated with this figure shows that there is an improvement with the correlations from Figure 4 to Figure 8. In Figure 4, the middle of the graph had a higher correlation than the ends of the graph that showed a slight dip in the data. In Figure 8, the only “dip” was on the left end of the graph with the lower kernel values while the middle of the graph towards the right end showed an increased linear trend with consistent slopes across all BLOSUM matrices.

DISCUSSION:

There were multiple design constraints for this project:

1. Comparison with the BLOSUM matrices (30, 62, and 90) which characterized the substitution ability of each pair of amino acids as found in nature.

2. Use of the molecular signature as a way to describe the structure of the amino acids. Since the molecule signature descriptor is a public resource, no funds were needed for this project.

3. The signature kernel equation that was used in order to repeat a previous study by Jean-Loup Faulon.

Through all three aspects of this Honors Project, it can be inferred that the molecular signature relies on both the polarity and the core backbone of the structure of the amino acid. Throughout the study, the linear relationship improved with each experiment which shows the importance of the height of the molecular signature and polarity of the side groups.

When analyzing the data, the trend of side groups did account for fluctuations in the data. In Table 1, some of the large and nonpolar amino acids had the lowest BLOSUM scores. This indicates polarity and size of the side group matters with regards to the substitution ability of amino acids for each other.

In each of the investigations, multiple BLOSUM matrices were used in order to prove consistency. According to Faulon, BLOSUM62 is the most used matrix since it is most consistently seen in nature.[1]

CONCLUSION:

This Honors Project focuses on furthering the study of using predictive mathematical models to determine if one amino acid can be substituted for another. Substitution of amino acids allows drugs to be developed using amino acids as substitutes for the targeted amino acid. A method to predict the success of this substitution mathematically would allow drugs to be developed faster and have a greater success rate when then used in human trials. This Honors Project focused on three aspects of this topic:

- Duplicating the work of Jean-Loup Faulon who used the molecular signature to investigate drug effectiveness
- Comparing the molecular signature kernels to three of the BLOSUM matrices (30, 62, and 90) to test the accuracy of the mathematical model
- Manipulating the kernel matrix in order to improve the relationship:
    - 1. Focusing on side groups
    - 2. Changing how the structure was represented in the matrix by increasing the initial height distance from the central atom (Height 1 and Height 2 included)

There were multiple design constraints for this project. The first was the comparison with the BLOSUM matrices (30,62, and 90) which characterized the substitution ability of each pair of amino acids as found in nature. The next constraint was the use of the commonly found molecular signatures on the internet. This Honors project did not have funding in order to use the paid for and more accurate molecular signature descriptor method. The last constraint was the usage of the signature kernel equation which was used in order to repeat a previous study by Jean-Loup Faulon.

The modified version of the molecular matrix improved the linear relationship R^2 value from 0.83 to 0.87. In Figure 7, the oscillation towards the middle of the graph indicates that the relationship is not very reliable. The modified matrix focuses on the side groups and therefore shows that it is not the only factor when finding similarity between two amino acids.

The height 2 version of the molecular matrix improved the linear relationship R^2 value from 0.83 to 0.93. This linear relationship is even higher than the modified version of the signature kernel values. In Figure 8, the only “dip” was on the left end of the graph with the lower kernel values while the middle of the graph towards the right end showed an increased linear trend with consistent slopes across all BLOSUM matrices. Although, there is a consistent upwards slope towards the larger kernel signature values at the right end of the graph which shows an improved relationship in comparison to Figure 4.

The modified matrix accounts for only the polarity and side groups which help explain why it is important to use an all-atom approach when finding similarities while using height 1. Height 2 focuses on an all- atom approach and shows an even better relationship. The substitution ability is complex and relies on many factors besides pure structure. While in this investigation, polarity was proved to not be the only source, it is still an important factor when describing the amino acid quantifiably.

RECOMMENDATIONS:

Due to a large improvement of the linear correlation R^2 value, it is apparent that height 2 was a better way to represent amino acid similarity. The height 1 matrix did not show the best linear relationship. A height greater than 2 could result with an even higher linear correlation value and could account for the three-dimensional connections of the amino acid. Further investigation should test different properties (size of amino acid and polarity) of heights greater than 2 in order to improve the molecular signature method.

Future investigations should also account for even higher heights and investigate different trends that mimic the trends found within the BLOSUM62 matrix. If further patterns are found within the BLOSUM62 matrix then it could be used to modify future signature kernel matrices and find even higher linear correlation than was found in this Honors Project. It is also recommended to keep in mind that multiple factors (different pH levels and intermolecular forces) occur at once and individual investigations may overlook the synergistic effect on the structure of the amino acid as seen in nature.

REFERENCES

[1] Faulon, Jean-Loup, Milind Misra, Ken Sale, and Rajat Sapra. "Genome Scale Enzyme–Metabolite and Drug–target Interaction Predictions Using the Signature Molecular Descriptor." *Bioinformatics* 9 Jan. 2008: 225-33.
URL: http://bioinformatics.oxfordjournals.org/content/24/2/225.full

[2] **“**The role of computational methods in the identification of bioactive compounds”. Meir Glick, Edgar Jacoby. *Current Opinion in Chemical Biology* 2011 *15* (4), 540-546.
DOI:10.1016/j.cbpa.2011.02.021.
<http://www.sciencedirect.com/science/article/pii/S1367593111000342>.

[3] “Designing Novel Polymers with Targeted Properties Using the Signature Molecular Descriptor”. W. Michael Brown,*,†,‖, Shawn Martin,†,‖, Mark D. Rintoul,† and, and Jean-Loup Faulon‡ *Journal of Chemical Information and Modeling* 2006 *46* (2), 826-835. DOI: 10.1021/ci0504521. < http://pubs.acs.org/doi/abs/10.1021/ci0504521>.

[4] "Alanine." *Analytical Wiki*. Wikia, 1 Jan. 2013. Web. 13 Apr. 2015.
<http://analytical.wikia.com/wiki/Alanine>.

[5] "Twenty Standard Amino Acids." *Exam 1 Review: Chapter 2 - Proteins and Enzymes*. Austin Peay State University, 18 Sept. 2010. Web. 2 Apr. 2015.
<http://apbrwww5.apsu.edu/thompsonj/Anatomy & Physiology/2010/2010 Exam Reviews/Exam 1 Review/Ch02 Protiens and Enzymes.htm>.

Appendix A

C++ Code used to calculate the Signature Kernels

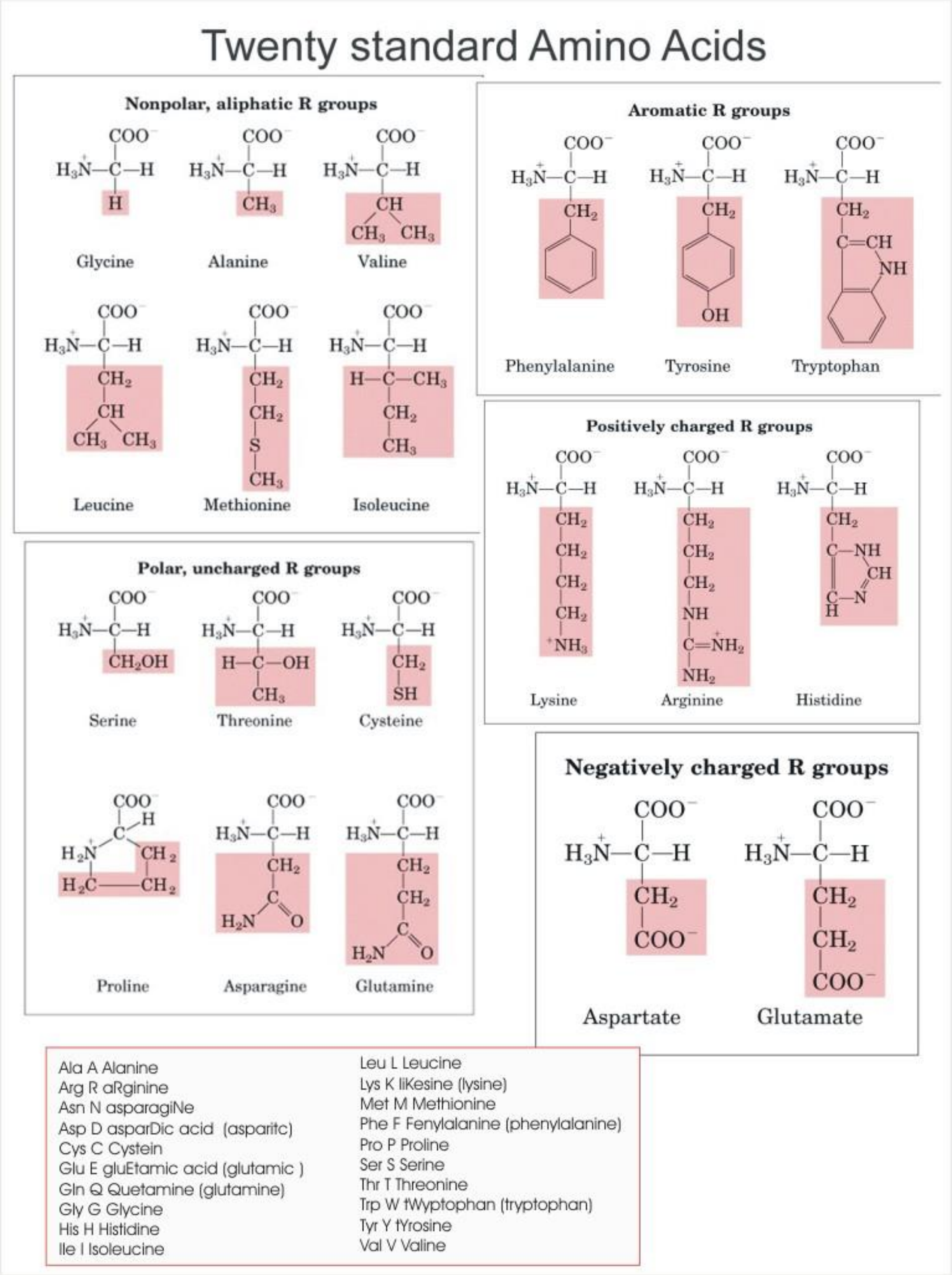
```
#include <iostream>
#include <fstream>
#include <cstdlib>
#include <cmath>
#include <string>
using namespace std;
int main()
{
   ifstream hc12;
   ifstream hc11;
   ofstream kernels("C:\\kernel.txt");
   double desmatrix1[21][34];
   double desmatrix2[21][34];

   int p, r, i, j, k, a, b, c, y, s;
   double prod;
   double prod1[21];
   double prod2;
   double prod3;
   double prod4;
   double prod5;
   double norm2[21][21];
   double norm1[21];
   double norm[21];
   double kernel[21][21];
   double kernel1[21][21];
   double kernel2[21][21];

   hc12.open("HCdesmatrix.txt");
   hc11.open("HCdesmatrix1.txt");

   if(!hc12)
   {cout << "File does not exist." << endl;}
   if(!hc11)
   {cout << "File does not exist." << endl;}

   for (p=1; p <21; p++){
   for(r=1; r < 14; r++)
   {hc12 >> desmatrix1[p][r];}}
   for (p=1; p <21; p++){
   for(r=1; r < 34; r++)
   {hc11 >> desmatrix2[p][r];}}

   for (i=1; i <21; i++){
      prod = 0;
       for (k=1; k <34; k++){
         hc12 >> desmatrix1[i][k];
         prod=prod+(desmatrix1[i][k]*desmatrix1[i][k]);
         norm1[i] = prod;

       for (s=1; s <21; s++){
```

```
    prod5 =0;

       for (y=1; y <34; y++){

         hc12 >> desmatrix1[i][y];

         hc12 >> desmatrix2[s][y];

         prod5 = prod5 + (desmatrix1[i][y]*desmatrix2[s][y]);

         norm2[i][s] = prod5;} }} }

   for (i=1; i <21; i++){ prod = 0;

       for (j=1; j <34; j++){

           prod=prod+(desmatrix1[i][j]*desmatrix1[i][j]);

           norm[i] = prod;}}


       for (p=1; p <21; p++){

       for(r=1; r < 21; r++)

   {hc12 >> desmatrix1[p][r];

      kernel1[p][r]= norm2[p][r] / sqrt((norm[p]*norm1[r]));

cout << kernel1[p][r] << " " << p << " " << r <<  endl;

kernels << kernel1[p][r] << " " << p << " " << r <<  endl;}}}

return 0; }
```

kernels – output matrix of the kernel data – to a datafile

Appendix B

1. Amino Acid structures and their side groups.



Ala A Alanine
Arg R aRginine
Asn N asparagiNe
Asp D asparDic acid (asparitc)
Cys C Cystein
Glu E gluEtamic acid (glutamic)
Gln Q Quetamine (glutamine)
Gly G Glycine
His H Histidine
Ile I Isoleucine

Leu L Leucine
Lys K liKesine (lysine)
Met M Methionine
Phe F Fenylalanine (phenylalanine)
Pro P Proline
Ser S Serine
Thr T Threonine
Trp W tWyptophan (tryptophan)
Tyr Y tYrosine
Val V Valine

Figure 7. Each Amino Acid has a side group with different properties. Aspartate is Aspartic Acid and Glutamate is Glutamic Acid. The amino acids shown were used in the investigation.[5]

2. Amino Acid Molecule Structure

Alanine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])-

2.0 [H]([N])--

4.0 [H]([C])-

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])--

0.0

Arginine

1.0 [O]([C][H])--

1.0 [O](=[C])--

2.0 [N]([C][H][H])--

1.0 [N]([C][C][H])

1.0 [N](=[C][H])

1.0 [H]([O])-

6.0 [H]([N])--

7.0 [H]([C])-

1.0 [C]([N][N]=[N])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][N])

1.0 [C]([C][C][H][N])--

2.0 [C]([C][C][H][H])

0.0

Asparagine

1.0 [O]([C][H])--

2.0 [O](=[C])--

2.0 [N]([C][H][H])--

1.0 [H]([O])-

4.0 [H]([N])--

3.0 [H]([C])-

1.0 [C]([C][O]=[O])--

1.0 [C]([C][N]=[O])

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Aspartic Acid

2.0 [O]([C][H])--

2.0 [O](=[C])--

1.0 [N]([C][H][H])--

2.0 [H]([O])

2.0 [H]([N])--

3.0 [H]([C])

2.0 [C]([C][O]=[O])--

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Cysteine

1.0 [S]([C][H])

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([S])

1.0 [H]([O])

2.0 [H]([N])--

3.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][S])

1.0 [C]([C][C][H][N])--

0.0

Glutamic Acid

2.0 [O]([C][H])--

2.0 [O](=[C])--

1.0 [N]([C][H][H])--

2.0 [H]([O])

2.0 [H]([N])--

5.0 [H]([C])

2.0 [C]([C][O]=[O])--

1.0 [C]([C][C][H][N])--

2.0 [C]([C][C][H][H])

0.0

Glutamine

1.0 [O]([C][H])--

2.0 [O](=[C])--

2.0 [N]([C][H][H])--

1.0 [H]([O])

4.0 [H]([N])--

5.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][N]=[O])

1.0 [C]([C][C][H][N])--

2.0 [C]([C][C][H][H])

0.0

Glycine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

2.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][N])--

0.0

Histidine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N](p[C]p[C][H])

1.0 [N](p[C]p[C])

1.0 [N]([C][H][H])--

1.0 [H]([O])

3.0 [H]([N])--

5.0 [H]([C])

1.0 [C](p[C][H]p[N])

1.0 [C]([H]p[N]p[N])

1.0 [C]([C]p[C]p[N])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Isoleucine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

10.0 [H]([C])

1.0 [C]([C][O]=[O])--

2.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

1.0 [C]([C][C][C][H])

0.0

Leucine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

10.0 [H]([C])

1.0 [C]([C][O]=[O])--

2.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

1.0 [C]([C][C][C][H])

0.0

Lysine

1.0 [O]([C][H])--

1.0 [O](=[C])--

2.0 [N]([C][H][H])--

1.0 [H]([O])

4.0 [H]([N])--

9.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][N])

1.0 [C]([C][C][H][N])--

3.0 [C]([C][C][H][H])

0.0

Methionine

1.0 [S]([C][C])

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

8.0 [H]([C])

1.0 [C]([H][H][H][S])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][S])

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Phenylalanine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

8.0 [H]([C])

5.0 [C](p[C]p[C][H])

1.0 [C]([C]p[C]p[C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Proline

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][C][H])

1.0 [H]([O])

1.0 [H]([N])--

7.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][N])

1.0 [C]([C][C][H][N])--

2.0 [C]([C][C][H][H])

0.0

Serine

2.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

2.0 [H]([O])

2.0 [H]([N])--

3.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][O])

1.0 [C]([C][C][H][N])--

0.0

Threonine

2.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

2.0 [H]([O])

2.0 [H]([N])--

5.0 [H]([C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][O])

1.0 [C]([C][C][H][N])--

0.0

Tryptophan

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N](p[C]p[C][H])

1.0 [N]([C][H][H])--

1.0 [H]([O])

3.0 [H]([N])--

8.0 [H]([C])

1.0 [C](p[C]p[C]p[N])

1.0 [C](p[C]p[C]p[C])

4.0 [C](p[C]p[C][H])

1.0 [C](p[C][H]p[N])

1.0 [C]([C]p[C]p[C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Tyrosine

2.0 [O]([C][H])--1

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

2.0 [H]([O])

2.0 [H]([N])--

7.0 [H]([C])

1.0 [C](p[C]p[C][O])

4.0 [C](p[C]p[C][H])

1.0 [C]([C]p[C]p[C])

1.0 [C]([C][O]=[O])--

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

0.0

Valine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

8.0 [H]([C])

1.0 [C]([C][O]=[O])--

2.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][C][H])

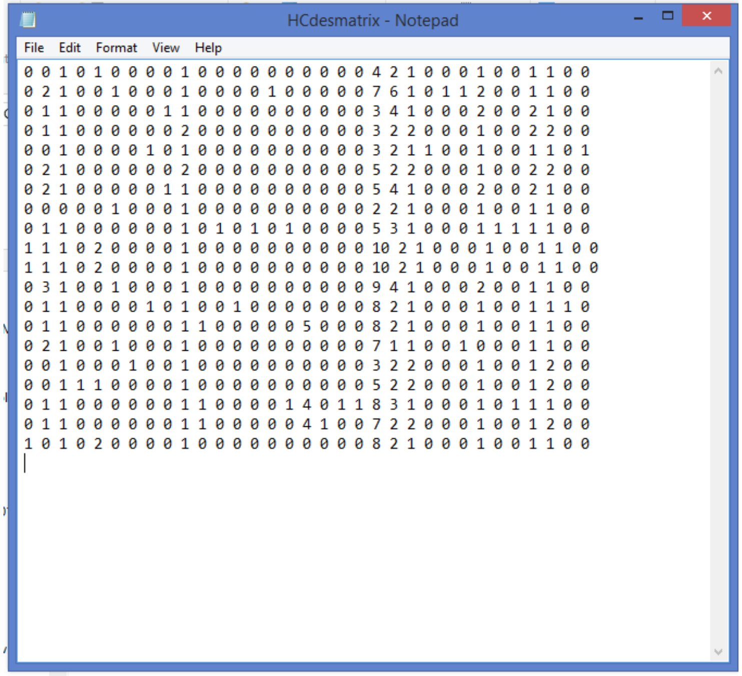
0.0

NOTE - The marked signatures are the common signatures that are removed or decreased to create the modified matrix.

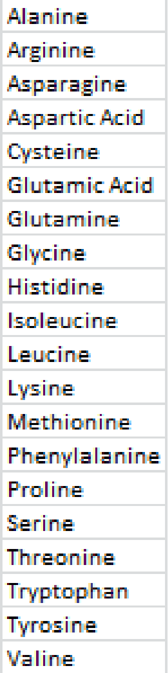
Appendix C

Input Matrices – Both Original Molecular Signature and Shortened Molecular Signature

Figure 8. The Matrix used for the Original Calculations. The Signatures describe the amino acids in alphabetical order as the second figure shows.

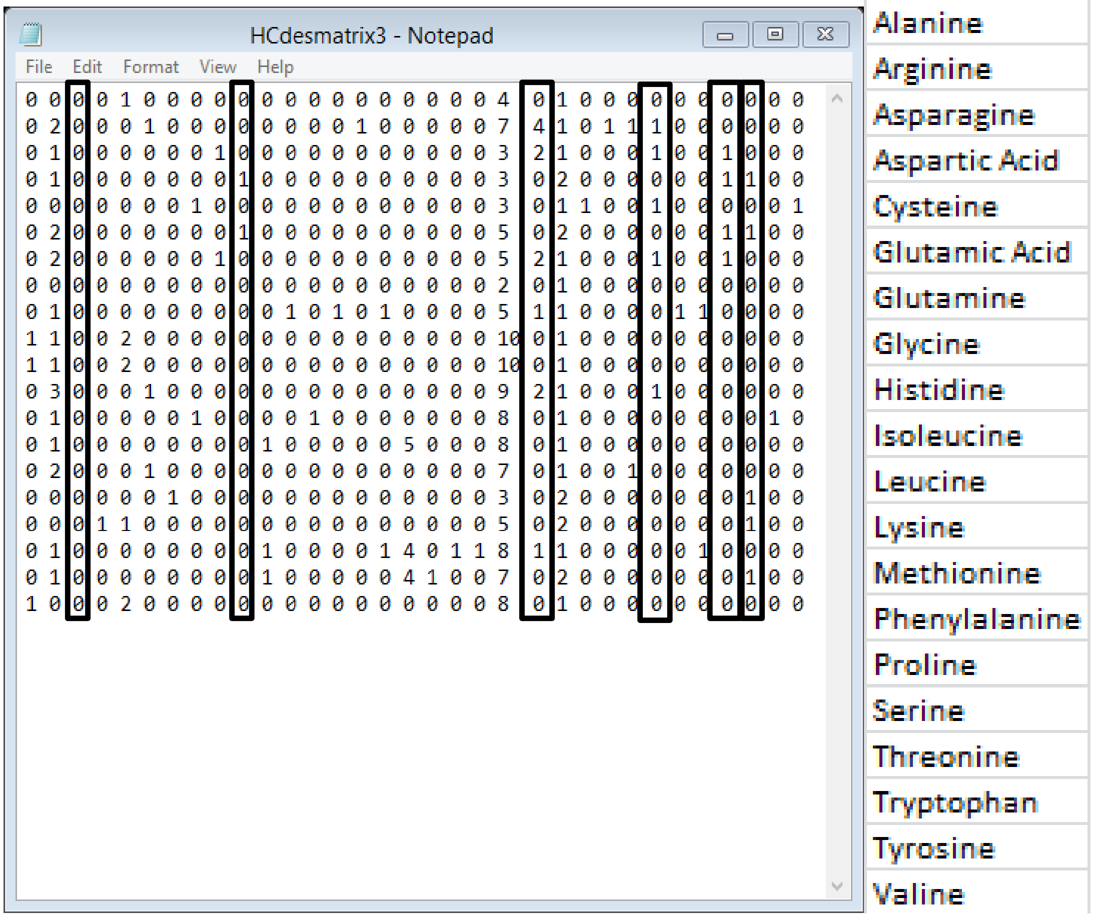


```
0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 4 2 1 0 0 0 1 0 0 1 1 0 0
0 2 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 7 6 1 0 1 1 2 0 0 1 1 0 0
0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 3 4 1 0 0 0 2 0 0 2 1 0 0
0 1 1 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 3 2 2 0 0 0 1 0 0 2 2 0 0
0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 3 2 1 1 0 0 1 0 0 1 1 0 1
0 2 1 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 5 2 2 0 0 0 1 0 0 2 2 0 0
0 2 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 5 4 1 0 0 0 2 0 0 2 1 0 0
0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 2 2 1 0 0 0 1 0 0 1 1 0 0
0 1 1 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 0 0 5 3 1 0 0 0 1 1 1 1 1 0 0
1 1 1 0 2 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 10 2 1 0 0 0 1 0 0 1 1 0 0
1 1 1 0 2 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 10 2 1 0 0 0 1 0 0 1 1 0 0
0 3 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 9 4 1 0 0 0 2 0 0 1 1 0 0
0 1 1 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 8 2 1 0 0 0 1 0 0 1 1 1 0
0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 5 0 0 0 8 2 1 0 0 0 1 0 0 1 1 0 0
0 2 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 7 1 1 0 0 1 0 0 0 1 1 0 0
0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 3 2 2 0 0 0 1 0 0 1 2 0 0
0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 5 2 2 0 0 0 1 0 0 1 2 0 0
0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1 4 0 1 1 8 3 1 0 0 0 1 0 1 1 1 0 0
0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 4 1 0 0 7 2 2 0 0 0 1 0 0 1 2 0 0
1 0 1 0 2 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 8 2 1 0 0 0 1 0 0 1 1 0 0
```



Alanine
Arginine
Asparagine
Aspartic Acid
Cysteine
Glutamic Acid
Glutamine
Glycine
Histidine
Isoleucine
Leucine
Lysine
Methionine
Phenylalanine
Proline
Serine
Threonine
Tryptophan
Tyrosine
Valine

Figure 9. The matrix used for the improved and modified version of the Molecular Signatures. The Signatures describe the amino acids in alphabetical order as the second figure shows.



```
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4  0 1 0 0 0 0 0 0 0 0 0 0
0 2 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 7  4 1 0 1 1 1 0 0 0 0 0 0
0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 3  2 1 0 0 0 1 0 0 1 0 0 0
0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 3  0 2 0 0 0 0 0 0 1 1 0 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 3  0 1 1 0 0 1 0 0 0 0 0 1
0 2 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 5  0 2 0 0 0 0 0 0 1 1 0 0
0 2 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 5  2 1 0 0 0 1 0 0 1 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2  0 1 0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 5  1 1 0 0 0 0 1 1 0 0 0 0
1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 10 0 1 0 0 0 0 0 0 0 0 0 0
1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 10 0 1 0 0 0 0 0 0 0 0 0 0
0 3 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9  2 1 0 0 0 1 0 0 0 0 0 0
0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 8  0 1 0 0 0 0 0 0 0 0 1 0
0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 5 0 0 0 8  0 1 0 0 0 0 0 0 0 0 0 0
0 2 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7  0 1 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 3  0 2 0 0 0 0 0 0 0 1 0 0
0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5  0 2 0 0 0 0 0 0 0 1 0 0
0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 4 0 1 1 8  1 1 0 0 0 0 0 1 0 0 0 0
0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 4 1 0 0 7  0 2 0 0 0 0 0 0 0 1 0 0
1 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8  0 1 0 0 0 0 0 0 0 0 0 0
```

Alanine
Arginine
Asparagine
Aspartic Acid
Cysteine
Glutamic Acid
Glutamine
Glycine
Histidine
Isoleucine
Leucine
Lysine
Methionine
Phenylalanine
Proline
Serine
Threonine
Tryptophan
Tyrosine
Valine

The modified matrix was created by finding six columns that were the same in each amino acid row. The modified columns were shown circled in Figure 9. The common number of similarities was found represented the backbone (carboxylic acid) of the amino acids. The common values were reduced or put as zero. The new matrix will focus more on the structure of the side chains and display polarity.

Appendix D

Example calculation for the molecular signature kernel.

Alanine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])—

2.0 [H]([N])--

4.0 [H]([C])-

1.0 [C]([C][O]=[O])--

1.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])--

0.0

σ(Aln) = (0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 2, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0)

$|\boldsymbol{\sigma}(\mathbf{Aln})|^2$ **= (1+1+1+16+4+1+1+1+1) = 27**

$$|\boldsymbol{\sigma}(\mathbf{Aln})| = \sqrt{\mathbf{27}}$$

Leucine

1.0 [O]([C][H])--

1.0 [O](=[C])--

1.0 [N]([C][H][H])--

1.0 [H]([O])

2.0 [H]([N])--

10.0 [H]([C])

1.0 [C]([C][O]=[O])--

2.0 [C]([C][H][H][H])

1.0 [C]([C][C][H][N])--

1.0 [C]([C][C][H][H])

1.0 [C]([C][C][C][H])

0.0

σ(Leu) = (1, 1, 1, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 10, 2, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0)

$|\boldsymbol{\sigma}(\mathbf{Leu})|^2$**=(1+1+1+4+1+100+4+1+1+1+1)= 116**

$$|\boldsymbol{\sigma}(\mathbf{Leu})| = \sqrt{\mathbf{116}}$$

$$\mathbf{k(Aln, Leu)} = \frac{\boldsymbol{\sigma}(\boldsymbol{Aln})\cdot\,\boldsymbol{\sigma}(\boldsymbol{Leu})}{|\boldsymbol{\sigma}(\mathbf{Aln})|\;|\boldsymbol{\sigma}(\mathbf{Leu})|} = \mathbf{0.92}$$