

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

ABSTRACT

Bio molecular geometry and physiochemical composition are frequently described and analysed by graphs and graphical methods. This way of describing mostly defines similarity measures of these structures. In this regard, graph kernels occupy a wide space in recent developments. In this paper, we propose a simple and a new graph kernel based on path length between each pair of vertices. The performance of this kernel shows a significant classification of accuracy.

Keywords: *Graph kernel, similarity measure, SSE of Protein, Path.*

Introduction

Starting from social network analysis through computational biology to neural network the data is modelled as graph and in many of these we handle a large scale of data. In computational biology, functional analysis of protein is the key research problem. The increased number of protein structures increases structure based prediction of protein function as well as classical sequence based prediction methods. As functional similarity of protein does not necessarily come along with sequence similarity [1], the structure based prediction methods have increased their own space. Here comes the measure of similarity between proteins based on their structures. When proteins are transformed into graphs the measure of similarity between proteins gets diverted as a measure of similarity between graphs. Some graph similarity measures are

- i. Using graph edit distance, similarity is measured with respect to the topology [20, 21].
- ii. Skew spectrum restricted to unlabeled graphs [22].
- iii. Graph let spectrum approach [23].

Apart from these, graph kernels - finding the best match between substructures of graphs has evolved into an emerging branch of learning on structured data. Kernel is

from statistical learning theory [1] and performs the following tasks (i) classification using support vector machines [24], (ii) regression [2], (iii) clustering [3] and (iv) principal component analysis [4].

Haussler [5], the first to define a unique way of designing kernels on structured objects and is known as R- Convolution kernel. The kernels that exist now can be viewed as a special case of the R- Convolution kernel. Existing kernels are the kernels on structured objects in graphs [6] and on graphs [7, 8].

D.VIJAYALAKSHMI

Assistant Professor,

Department of Mathematics,

Sri Chandrasekharendra Saraswathi

Viswa Mahavidyalaya, Kanchipuram -

631 561, Tamil Nadu, INDIA

guruviji97@gmail.com

K.SRINIVASA RAO

Professor,

Department of Mathematics,

Sri Chandrasekharendra

Saraswathi Viswa Mahavidyalaya

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

Graph kernels defined can be categorized into

- i. Graph kernels based on walks [8, 9]. This determines the number of matching pairs of random walks in given graphs.
- ii. Graph kernels based on path [10]. This determines the number of pairs of shortest paths having same initial and final vertex and of same length.
- iii. Graph kernels based on limited size sub graphs [12, 13]. This determines the number of sub graphs of size k, k can be size 3, 4, 5 of all types.
- iv. Graph kernels based on sub tree patterns [13, 14]. This determines the number of pairs of matching substructure in sub tree pattern on comparing the corresponding vertices of two graphs.

Apart from these, we have several graph kernels and they are

- i. Computation of random walk kernel using dynamic programming, the cost is of considering walks of fixed size [15].
- ii. Extending the marginalized graph kernels [8] by relabeling the vertices of a graph using Morgan index [16].
- iii. Determining the number of identical pairs of rooted sub graphs containing vertices up to a certain distance from root and the roots are located at a certain distance from each other in two graphs [17].
- iv. Refined version of Ramon – Gartner kernel [18, 19].

This paper is designed in the following way. In the first section method of graph construction is narrated, followed by path kernel definition. The steps while applying

this method is also explained in detailed in the first section. Next is the result section where results obtained by path kernel method is given with the results obtained by smith waterman method. At last the conclusion saying the positive and negative points of path kernel method.

Method

The protein graph is constructed by considering the secondary structure elements as vertices. The vertices are named according to their structures in sequential order. Here we consider beta strands, helices, beta turns. i.e., the vertex corresponding to first strand is named as s1 and second one as s2 and so on. Average of 3-D co-ordinates of central carbon atom of each amino acid in SSE is calculated and named as centroid. To decide the edges, the distance between centroid of a vertex to the remaining vertices is calculated. Two vertices of least distance from the vertex are connected by edges. This method is explained in [25]. Next comes the definition of kernel based on path length.

$$P[K(G, G')] = \sum_{h=0}^2 K_h(G, G')$$

$$i.eP[K(G, G')] = K_0(G, G') + K_1(G, G') + K_2(G, G')$$

Where

$\delta(l(v), l(v'))$ is an indicator function and is defined by

$$\delta(l(v), l(v')) = \begin{cases} 1 & \text{if the arguments are equal} \\ 0 & \text{otherwise} \end{cases}$$

$$K_0(G, G') = \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v'))$$

this counts the number of common labelled vertices.

$$K_1(G, G') = \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \cdot N_1(v, v')$$

this counts the number of common labelled

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

vertices at a distance of path length one. $N_1(v, v')$ is set of common neighbourhood vertices at distance of path length 1 from v, v' vertices.

$$K_2(G, G') = \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')). N_2(v, v')$$

this counts the number of common labelled vertices at a distance of path length two. $N_2(v, v')$ is set of common neighbourhood vertices at distance of path length 2 from v, v' vertices.

Here we use normalised kernel to measure similarity. The normalised kernel is defined as

$$P[K(G, G')] = \frac{P[K(G, G')]}{\sqrt{P[K(G, G)] \cdot P[K(G', G')]}}$$

In this section we apply this path kernel to graphs of same number of vertices. This kernel function is applied to seven vertices graphs and six vertices graphs.

This kernel function is symmetric and positive definite and the same is verified in the following steps.

Symmetric condition

Let us consider

$$P[K(G, G^1)] = \sum_{h=0}^2 K_h(G, G^1)$$

This function counts the number of common labelled vertices for each pair of corresponding vertices of two graphs. In this the first element in the pair is from the graph G and the second element from the graph G^1 . Simply this function counts the common labelled vertices at path length zero, one and two. Now we consider


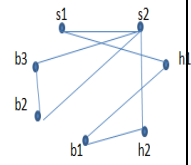
$$P[K(G^1, G)] = \sum_{h=0}^2 K_h(G^1, G)$$

This function also counts the set of similar labelled vertices for each pair of vertices from two graphs at distance of path length zero, one, two. Only thing that differs is the first element is from the graph G^1 and second element is from graph G .


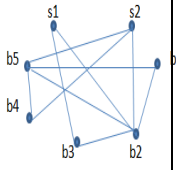

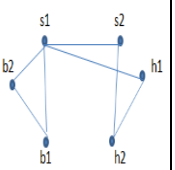

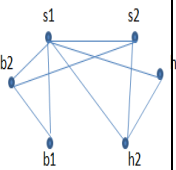

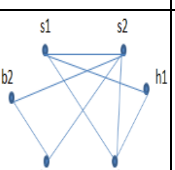

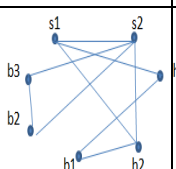

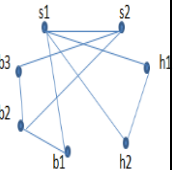

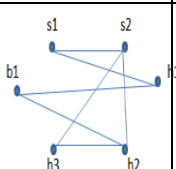

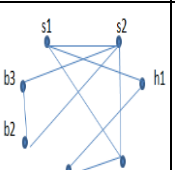

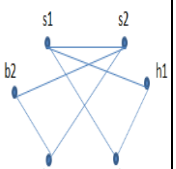

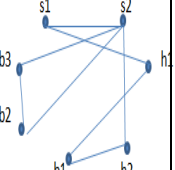

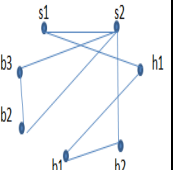
Both yields the same result as both is the cardinality of the set of common labelled vertices at distance zero, one, two calculated for each pair of vertices from the two graphs. Therefore $P[K(G, G')] = P[K(G^1, G)]$. Hence the function is symmetric

This function is definitely positive definite because, firstly the graphs are labelled in similar manner. Secondly the first step in this kernel calculation is comparing the original labels of the graphs and this enables the function to secure positive value. This is because a protein structure can have all the secondary structure element or at least two secondary structures. This fact reveals that the path kernel is positive definite.

The following table gives the details of proteins and the graph obtained for each protein. As we detail with SSE the structure of protein is also given.

S.no	Protein Structure	Graph for Protein	Protein name
1			Crambin mixed sequence form at 160 k. Protein/water substates

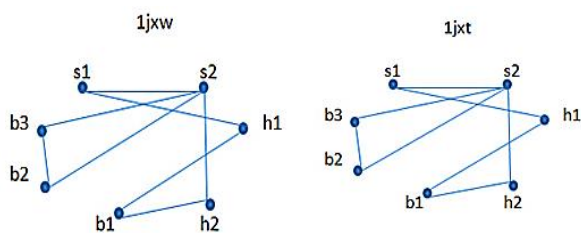
PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

2	1nbl 		NMR Structure of Hellethionin D	8	2v9b 		X-ray structure of viscotoxin b2 from viscum album
3	1bhp 		Structure of beta-purothionin at room temperature and 1.7 angstroms resolution	9	2plh 		Structure of alpha-1-purothionin at room temperature and 2.8 angstroms resolution
4	1jxu 		Crambin mixed sequence form at 240 k. Protein/water substates	10	1edo 		NMR structural determination of viscotoxin A3 from Viscum album L.
5	2fd7 		X-ray Crystal Structure of Chemically Synthesized Crambin	11	1ccn 		Direct noe refinement of crambin from 2d NMR data using a slow-cooling annealing protocol
6	1wuw 		Crystal Structure of beta-hordothionin	12	1jxy 		Crambin mixed sequence form at 220 k. Protein/water substates
7	1jxw 		Crambin mixed sequence form at 180 k. Protein/water substates				

Table(i) Gives the protein details

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

Similarity/dissimilarity study of protein 1jxw and 1jxt using Path kernel



Protein graphs of 1jxw and 1jxt proteins represented by secondary structure elements

Step 1:

Initially we count the similar vertex labels from the two graphs

$$\begin{aligned}
 K_0(G, G') &= \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \\
 &= \delta(s1, s1) + \delta(s2, s2) + \delta(h1, h1) + \delta(h2, h2) + \\
 &\quad \delta(b1, b1) + \delta(b2, b2) + \delta(b3, b3) \\
 &= 1 + 1 + 1 + 1 + 1 + 1 + 1 \\
 &= 7
 \end{aligned}$$

Step 2:

We count the similar labelled vertices for each pair of corresponding vertices from the two graphs at a distance of path length one.

$$\begin{aligned}
 K_1(G, G') &= \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \cdot N_1(v, v') \\
 &= \delta(s1, s1) \cdot 2 + \delta(s2, s2) \cdot 4 + \delta(h1, h1) \cdot 2 + \delta(h2, h2) \cdot 2 + \\
 &\quad \delta(b1, b1) \cdot 2 + \delta(b2, b2) \cdot 2 + \delta(b3, b3) \cdot 2 \\
 &= 1.2 + 1.4 + 1.2 + 1.2 + 1.2 + 1.2 + 1.2 \\
 &= 16
 \end{aligned}$$

Step 3:

In this step we count common labelled vertices for each pair of corresponding vertices from the two graphs at a distance of path length two.

$$\begin{aligned}
 K_2(G, G') &= \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \cdot N_2(v, v') \\
 &= \delta(s1, s1) \cdot 4 + \delta(s2, s2) \cdot 4 + \delta(h1, h1) \cdot 2 + \\
 &\quad \delta(h2, h2) \cdot 4 + \delta(b1, b1) \cdot 2 + \delta(b2, b2) \cdot 4 + \\
 &\quad \delta(b3, b3) \cdot 4 \\
 &= 1.4 + 1.4 + 1.2 + 1.4 + 1.2 + 1.4 + 1.4 \\
 &= 24
 \end{aligned}$$

Step 4:

The path kernel

$$P[K(G, G')] = 7 + 16 + 24 = 47$$

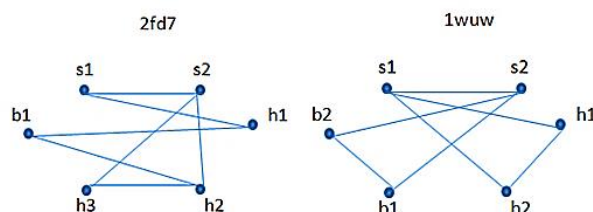
Step 5:

The normalised path kernel

$$P[K(G, G')] = \frac{47}{\sqrt{47 \cdot 47}} = 1$$

The protein graphs are 100% similar this implies the proteins 1jxt and 1jxw are 100% similar.

Similarity/Dissimilarity study of protein 2fd7 and 1wuw using Path kernel



Protein graphs of 2fd7 and 1wuw proteins represented by secondary structure elements

Step 1: Initially we count the similar vertex labels from the two graphs

$$\begin{aligned}
 K_0(G, G') &= \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \\
 &= \delta(s1, s1) + \delta(s2, s2) + \delta(h1, h1) + \delta(h2, h2) + \delta(b1, b1) \\
 &= 1 + 1 + 1 + 1 + 1 \\
 &= 5
 \end{aligned}$$

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

Step 2:

We count the similar labelled vertices for each pair of corresponding vertices from the two graphs at a distance of path length one.

$$K_1(G, G') = \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \cdot N_1(v, v')$$

$$= \delta(s1, s1) \cdot 2 + \delta(s2, s2) \cdot 1 + \delta(h1, h1) \cdot 1 + \delta(h2, h2) \cdot 0 + \delta(b1, b1) \cdot 0$$

$$= 1 \cdot 2 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 0$$

$$= 4$$

Step 3:

In this step we count common labelled vertices for each pair of corresponding vertices from the two graphs at a distance of path length two.

$$K_2(G, G') = \sum_{v \in G} \sum_{v' \in G'} \delta(l(v), l(v')) \cdot N_2(v, v')$$

$$= \delta(s1, s1) \cdot 2 + \delta(s2, s2) \cdot 3 + \delta(h1, h1) \cdot 2 + \delta(h2, h2) \cdot 3 + \delta(b1, b1) \cdot 2$$

$$= 1 \cdot 2 + 1 \cdot 3 + 1 \cdot 2 + 1 \cdot 3 + 1 \cdot 2$$

$$= 12$$

Step 4:

The path kernel

$$P[K(G, G')] = 5 + 4 + 12 = 21$$

Step 5:

The normalised path kernel

$$P[K(G, G')] = \frac{21}{\sqrt{40.40}} = 0.525$$

The protein graphs are 52% similar this implies the proteins 2fd7 and 1wuw are 52% similar.

Result

Results of Proteins with Seven SSE

	ljxt	ljxw	lnbl	led0	lccn	ljxu
ljxy	1 1	1 1	0.364 0.34- 0.53	0.695 0.48- 0.62	0.942 0.95- 0.97	0.94 1
ljxt	-	1 1	0.364 0.34- 0.53	0.695 0.48- 0.62	0.942 0.95- 0.97	0.94 1
ljxw	-	-	0.364 0.34- 0.53	0.695 0.48- 0.62	0.942 0.95- 0.97	0.94 1
lnbl	-	-	-	0.357 0.55- 0.66	0.343 0.32- 0.51	0.327 0.34- 0.51
led0	-	-	-	-	0.729 0.65	0.729 0.68
lccn	-	-	-	-	-	1 0.96

Table(ii) Results obtained by **Path kernel method** and Smith Waterman method for 7 vertices.

In the above table, the value at the top is the result obtained by our method and the value at the bottom refers the result by Smith Waterman method. The value below narrates identical amino acids to similar amino acids in the protein sequences under comparison. In the above table the results corresponding to the protein lnbl, the results is around the percentage of identical amino acid in the protein sequence, this is because the similarity is measure based on SSEs , in lnbl there are two beta strands and five beta turn structures whereas the remaining proteins have all type of structures-beta strands, helices, beta turn structures.

Results of Protein with Six SSE

	2v9b	2plh	2fd7	1wuw
1bhp	0.839 0.69	0.89 0.88- 0.95	0.495 0.37- 0.58	0.775 0.80

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

2v9b	-	0.695 0.47- 0.65	0.487 0.46- 0.65	0.71 0.67
2plh	-	-	0.519 0.57	0.904 0.86
2fd7	-	-	-	0.525 0.53

Table(iii) Results obtained by **Path kernel method** and Smith Waterman method for 6 vertices.

While calculating the correlation between these values we get correlation coefficient as 0.911

Conclusion

Graph kernel defined in this part is based on the path length.

1. Here we consider set of common neighbourhood vertices at the distance of path length 0, 1 and 2.
2. This method is simple and can be extended ie. We can consider the set of common neighbourhood vertices at distance of path length 3 or more. And in the place of SSE we can consider each amino acid and its carbon co-ordinates to construct the graph.
3. Identifying the vertices based on the corresponding secondary structure shows its uniqueness.

The accuracy in result is reached using this kernel. But this kernel has a restriction that it shows its efficiency in reaching accuracy if the graphs have the same number of vertices. Path kernel method has an additional constraint and that is the shape of secondary structures. This makes the method to give the accurate percentage of identical part of amino acids in the two

protein sequences under consideration and increases the efficiency of the method.

References

1. B.Schlkopf and A.J.Smola (2002), *Learning with kernels*, MIT Press.
2. H.Drucker, C.J.C.Burges, L.Kaufman, A.Smola and V.Vapnik, (1997) *Support vector regression machine*. In M.C.Mozer, M.I.Jordan and T.Petsche, editor, *Advances in neural information processing systems 9*, Cambridge, MA, MIT Press 155-161.
3. A.Ben-Hur, D.Horn, H.Siegelmann and V.Vapnik, (2001), *A support vector method for hierarchical clustering*. In T.K.Leen, T.G.Dietterich and V.Tresp, editor, *Advances in neural information processing systems 13*, MIT Press, 367-373.
4. B.Schlkopf and A.J.Smola and K.R.Muller, (1999), *Kernel principal component analysis*. In B.Schlkopf and A.J.Smola, C.J.C.Burges, editors, *Advances in kernel methods support vector learning*, Cambridge, MA, MIT Press, 327-352.
5. D.Haussler, (1999), *Convolution kernels on discrete structures*, Computer science department, US Santa Cruz, Technical report UCSC-CRL-99-10.
6. R.S.Kondor and J. Lafferty, (2002) *Diffusion kernels on graphs and other discrete structures*. In *proceedings of the ICML*.
7. T.Gartner, (2002) *Exponential and geometric kernels for graphs*, In *NIPS workshop on unreal data*, Volume principles of modeling non vectorial data.
8. H.Kashima, K.Tsuda and A.Inokuchi, (2003), *Marginalized kernels between labeled graphs*. In *proceedings of the*

PATH KERNEL AND PROTEIN SIMILARITY/ DISSIMILARITY STUDY

- 20th international conference on machine learning [ICML] Washington DC, US.
9. T.Gartner, P.A. Flach and S.Wrobel, (2003), On graph kernel Hardness results and efficient alternatives. In B. Scholkopf and M.Warmuth, editor, 16th Annual conference on computational learning theory and 7th kernel workshop, COLT, springer).
 10. K.M.Borgardt and H.P.Kriegel, (2005), Shortest- path kernels on graphs. In proceedings of the international conference on Data mining, 74-81.
 11. T.Horvath, T.Gartner and S.Wrobel, (2004), Cyclic pattern kernels for predictive graph mining, In the proceedings of the international Conference on Knowledge Discovery and Data mining, 158-167.
 12. N.Sheravashidze, S.V.N.Vishwanathan, T.Petri, K.Melhorn and K.M.Borgwardt, (2009), Efficient graphlet kernels for large graph comparisons. In artificial intelligence and statistics.
 13. P.Mahe, J.P.Vert, (2006) Graph kernels based on tree patterns for molecules, q-bio/0609024.
 14. J.Ramon and T.Gartner, (2003) Expressivity versus efficiency of graph kernels, technical report, First International workshop on Mining Graphs, Trees and Sequences.
 15. Z.Harchaoui and F.Bach. (2007), Image classification with segmentation graph kernels. In proceedings of the IEEE conference on computer vision and pattern recognition.
 16. H.L.Morgan, The generation of unique machine description for chemical structures –a technique developed at chemical abstract service. *Journal of chemical documentations* 5(2):107-113.
 17. F.Costa and K.De.Grave (2010) Fast neighborhood sub graph pair wise distance kernel. In proceeding of international conference on machine learning, pages, 255-262.
 18. P.Mahe, J.P.Vert, (2009), Graph kernels based on tree patterns for molecules, *Machine Learning*, 75(1), 3- 35.
 19. F.R.Bach, (2008), Graph kernels between point clouds, In proceeding of the international conference on machine learning, 25-32.
 20. H.Bunke and G. Allermann. (1983), In exact graph matching for structural pattern recognition. *Pattern recognition letters*, 1: 245-253.
 21. M.Neuhaus and H.Bunke, (2005), Self-organizing maps for learning the edit costs in graph matching *IEEE Transactions on systems, Man and cybernetics, part B*, 35(3), 503-514.
 22. I.R.Kondor and K.M.Borgwardt (2008), The skew spectrum of graphs. In proceeding of the international conference on machine learning, 496-503.
 23. I.R.Kondor, N.Sheravashidze, and K.M.Borgwardt, (2009), The graphlet spectrum. In proceeding of the international conference on machine learning, 529-536.
 24. V.Vapnik, (1998), *Statistical learning theory*, Wiley New York.