IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Representation-dimensionality Trade-off in Biological Sequence-based Inference

Bahman Asadi
School of Electronics and Computer Science,
University of Southampton,
Southampton, UK

Mahesan Niranjan
School of Electronics and Computer Science,
University of Southampton,
Southampton, UK
mn@ecs.soton.ac.uk

***Abstract*—Statistical inference from the analysis of biological sequences is widely used in the prediction of structure and biochemical functions of newly found macromolecules. For the application of machine learning methodologies such as kernel methods and artificial neural networks for such inference, variable length sequence data is often embedded in a finite dimensional real-valued space. The corresponding embedding dimensions are often high, leading to technical difficulties centred around the statistical concept of the *curse of dimensionality*. We demonstrate a trade-off between fidelity of representation of amino acids of proteins and the resulting dimensionality of the embedding space. Clustering chemically similar amino acids, thereby reducing the alphabet size, reduces the accuracy in their variation, but achieves a reduction in the corresponding feature space. We show this trade-off in three different problems of statistical inference, namely, protein-protein interaction, remote homology and secondary structure prediction. We show that in the reduced space performance often improves similar to what is seen in "diminishing returns" type reward-effort curves. We find alphabet reduction schemes taken from the literature, which are based on some biochemical rationale, perform significantly better than arbitrary random clustering of the alphabets. Statistical feature selection from the full 20 amino acid representation is not competitive with any of these. Dimensionality of representation has an important role when mapping sequence data onto fixed dimensions of an Euclidean space. This work shows that dimensionality reduction based on compressing the amino acid alphabet improves inference performance in two widely studied problems and degrades gracefully in the third. Alphabet reduction, which has a principled biochemical basis, is shown to be superior to feature selection which is purely a statistical exercise.**

***Index Terms*—Bioinformatics, Biological Sequence Analysis, Curse of Dimensionality.**

I. Introduction

DATA-DRIVEN inferences from protein sequences continues to be an important problem in bioinformatics and computational biology. This is because, data relating to protein sequences is growing at a very rapid pace when compared to data on three dimensional structures or experimentally verified *in vivo* biochemical functional annotations. Because our understanding of sequence to structure to functional relationships remains poor, computational learning methods become helpful in learning relationships from existing data and generalizing to make predictions about unseen data. Predicting secondary structures [1]–[4], inferring homologous relationships [5] and predicting protein-protein interactions [6] are examples of this.

At the core of developing data-driven methodologies for inference, is the representation of protein sequences, or the similarity between pairs of sequences, in a vector space where certain mathematical properties hold. Such embeddings, mapping symbolic data (chains of varying lengths formed by selecting from 20 amino acids), into a continuous mathematical space, enables the application of powerful machine learning methods that blend a rich mixture of function approximation / interpolation methods and statistical parameter estimation formulations. As protein sequences are of varying lengths, different coding strategies have been used to map them into finite dimensional spaces.

Often the dimensionality of the embedding space is high: a secondary structure prediction formulation using one out of N binary coding and a window size of 13 has 260 dimensions while a string kernel using triplet occurrence frequencies is of dimension 8000. Statistical pattern recognition in high dimensions is notoriously difficult. If we were to characterize the probability density of a class of data in high dimensional spaces, then it is known that to maintain the same precision, one needs exponentially more data with increasing dimensions, a phenomenon known as the *curse of dimensionality* [7]–[9]. In this perspective, regression and classification problems are considered easier than density modelling [10]. Nevertheless, high dimensions still make inferences difficult and a plethora of methods including projections onto principal subspaces, feature selection [11], [12] and sparsity inducing regularization have been explored to circumvent this particular issue.

A particular way to achieve dimensionality reduction, which we explore in this paper, is to compress the alphabet. Amino acids with similar chemical properties can be merged and be represented by one symbol for the purpose of encoding. Such alphabet reduction schemes have been explored by a number of authors (*e.g.* [13]–[19]).

Peterson *et al.* [13] explore a wide range of alphabet reduction schemes in sequence alignment-based fold recognition tasks. They use the DALI structural alignment database to obtain ground truth of alignment and carry out pairwise sequence alignments using 150 different alphabet reduction schemes [20]. The probabilistic interpretation of substitution costs is used to re-scale the cost matrices in a reduced alphabet scheme [21] . They show that reduced alphabets give improved performance in fold recognition.

978-1-7281-2009-6/$31.00 paper N-20023.pdf

An information theoretic design of reduced alphabet representation is pursued in Solis and Rackovsky's work [22]. Here, mutual information between sequences and local structures is used to seek a data compression (by alphabet reduction), which can be used to retrieve folds. The focus is on backbone structural information and the work shows that reducing the alphabet still retains the ability to identify the target structure of interest. Andersen and Brunak [23] present an elegant structure preserving alphabet reduction scheme which sequentially clusters down the alphabet by seeking to maintain structure retrieval accuracy across 650 non-sequence similar chains and demonstrate the efficiency of the reduction scheme in a phylogenetic reconstruction problem.

In this paper, we consider two inference problems based on protein sequences: the prediction of protein-protein interaction and the prediction of remote homologies, again based on sequence information alone. We also consider a small secondary structure prediction problem of classifying α helices using a small moving window of amino acids. We use these problems to demonstrate a compromise that exists between the resolution of representation by using a greater number of amino acid symbols and the resulting loss of statistical learning accuracy induced by expanded dimensionality. Where the dimensionality is high, as in the case of interaction prediction and homology detection, our results show that at reduced alphabet sizes a performance gain can be obtained, which then disappears if the quantization is made too coarse. In the case of the simple secondary structure classification problem, where we deliberately select a small dimensional representation to start with, the advantage obtained by alphabet reduction is not significant, yet the degradation in accuracy is graceful down to an alphabet size of 15 (rather than being sharp).

II. Methods

A. Dataset

Protein-protein Interaction:: For predicting protein-protein interactions, we downloaded the dataset used constructed by [24]. This is a state-of-the-art dataset constructed to carry out a critical appraisal on the performance of sequence based interaction prediction methods. It pays particular attention to balancing the different classes (interacting and non-interacting). Interactions are taken from experimental evidence and negative examples are set as complements of interacting proteins.

1) Homology Detection:: In order to analysing protein according to their structures and sequences, we used the Structural Classification of Protein (SCOP) database. In this database, proteins are placed in a hierarchical structure of classes, folds, superfamilies and families based on their evolutionary relationships and structural and functional similarities. Biologist have gathered these information through manual experiments which is why the database is considered to be very reliable. As Fig. 1A simple representation of the SCOP hierarchy and how the training and testing data sets have been split upfigure.1 shows, the highest level of the hierarchy is class in which the secondary structure similarities is the criteria for belonging to a class. The second level of the hierarchy is fold where the major secondary structure has the same arrangement with the same topological connections. The third level belongs to superfamilies. In this level, proteins have probable common evolutionary origin. In this level the sequence identities are low but the structural and functional features indicate a probable common evolutionary origin. And the lowest level would belong to families where proteins have clear evolutionarily relationships. For experimentation, we followed precisely the schemes adopted by Jaakkola *et al.* [26] and Wieser and Niranjan [5].

2) Secondary Structure Prediction: For secondary structure prediction, we used the `CB513` dataset developed by Cuff and Barton [27], which is a non-homologous data set containing 513 proteins.

B. Amino Acid's Alphabet Reduction

Amino acids have different characteristics which could be used for grouping them together. Peterson has explored the effects of amino acid reduction, in the distance of pairwise protein alignments. The results obtained from the reduced alphabets were close or even better in some cases, compared to the full length proteins [13].

These alphabets have been gathered from different research which have been used over time and are based on some measure of relative similarity. There are 14 different schemes overall which have different group sizes. Table IExample of reduced amino acid alphabet schemes, with alphabet sizes in the range 2 to 10 used by Peterson *et al.* [13]. Altogether we used a total of 46 different schemes taken from various papers in literaturetable.1 is an example for showing how the grouping is shown for the DSSP [22]. For a list of all the schemes refer to the original paper [13].

C. Bag of triplets representation

Mapping variable length sequence data onto fixed dimensions was achieved by frequencies of triplet occurrences, similar to the "bag of words" representation used in text analysis [28], and used in biological sequence representation by several authors including Shen *et al.* [6].

D. Sliding Window Encoding

The α helix prediction problem was formulated as a classification problem with a sliding window of 13 amino acids and an input encoding of one in N, following the framework of Qian and Sejnowski [1]. This is a simple coding scheme, and more successful secondary structure prediction approaches would align proteins in similar structural families and obtain a position specific probability distribution over the 20 amino acids. This gives a smoother representation to train classifiers on, and much of the improvement in secondary structure prediction over Qian and Sejnowski [1]'s initial formulation is attributed to such representation. However, this enhanced representation cannot be used in comparing amino acid reduction because the alignment and probability distributions have been computed with respect to a probability model (translated into scoring matrices). It is not straightforward to append

TABLE I
EXAMPLE OF REDUCED AMINO ACID ALPHABET SCHEMES, WITH ALPHABET SIZES IN THE RANGE 2 TO 10 USED BY PETERSON *et al.* [13]. ALTOGETHER WE USED A TOTAL OF 46 DIFFERENT SCHEMES TAKEN FROM VARIOUS PAPERS IN LITERATURE.

Alphabet size	Cluster									
	1	2	3	4	5	6	7	8	9	10
2	ACEFH IKLMQ RVWY	DGN PST								
3	AEHK QR	CFILM VWY	DGN PST							
4	AEHK QR	CFILM VWY	DNST	GP						
7	EKQR	FIV	LMWY	ACH	ST	DN	GP			
8	EKQR	IV	LWY	AM	CF	HT	DNS	GP		
9	EKQR	IV	L	F	AMW	CY	HT	DNS	GP	
10	EKQR	IV	LY	F	AM	W	HT	C	DNS	GP

an alphabet compression scheme onto these position specific distributions. However, since the objective of this work is to illustrate the effect of representation-dimensionality trade-off, rather than build a state-of-the-art secondary structure prediction system, it suffices to use the binary encoding scheme of [1].

E. Support Vector Machines

We used support vector machines as classifiers in all three of the inference problems. SVMs are kernel methods designed to induce generalization to unseen data by maximising the margin between a class boundary and training data that fall close to the boundary. A particular attraction of SVMs is that in modelling the boundary, it avoids representing the probability density of data, which is advantageous in high dimensions [10]. For implementation of SVM, we used the `LIBSVM` package [29]. Following previous authors, we used a Gaussian kernel in the protein-protein interaction problem, the default linear kernel for remote homology detection, and a Gaussian radial basis functions kernel for the α helix prediction problem.

F. Feature Selection

Feature selection was achieved by filtering, similar in concept to Golub *et al.* [30]. Following Guyon *et al.* [31], we used Pearson correlation coefficient to rank the features and picked the top few from the 8000 size feature space. The number of features retained was the same as the size of the feature space obtained with smaller alphabet sizes *i.e.* to compare with an alphabet size of 8, we would select $8 \times 8 \times 8$ features from the 8000 etc.

III. RESULTS AND DISCUSSION

A. Protein-Protein Interaction

Fig. 2Comparison of protein-protein interactions with different amino acid alphabet reduction schemes, showing the trade-off between dimensionality and resolution of representationfigure.2 shows prediction performances, quantified in terms of areas under receiver operating characteristic curves (AUC), for different combinations of amino acid alphabets. The AUC for the full 20 alphabet representation, which has an encoding dimension of 8000, is 0.75, matching the results quoted by Yu *et al.* [24]. This confirms similarity in experimental settings between our implementation and the previous work of Yu *et al.* [24]. The error bars, shown as boxplots in Fig. 2Comparison of protein-protein interactions with different amino acid alphabet reduction schemes, showing the trade-off between dimensionality and resolution of representationfigure.2, arise from the repertoire of different alphabet codes available in the literature. We note superior prediction performance for alphabet sizes around $7 \rightarrow 20$, significantly higher than at either end of the alphabet size scale. Surprisingly, even at very low resolution representations (alphabet size 3), reasonable prediction accuracies are obtained (median AUC 0.80), though the variation across the different coding schemes available is high.

As a sanity check, we also compared the performances of

Fig. 1. A simple representation of the SCOP hierarchy and how the training and testing data sets have been split up.

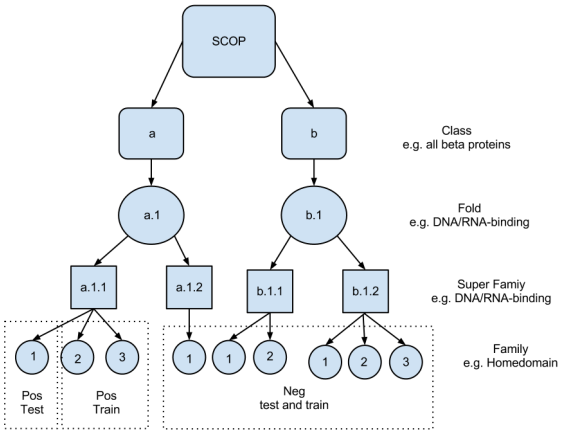
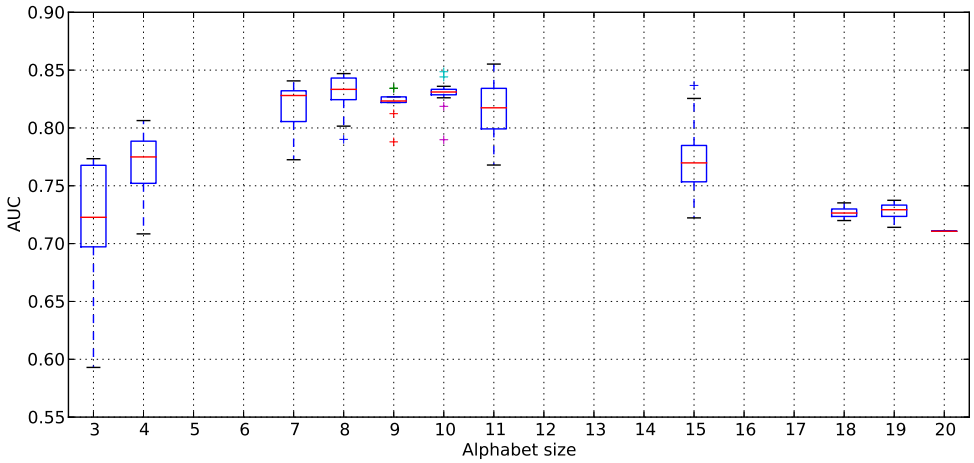
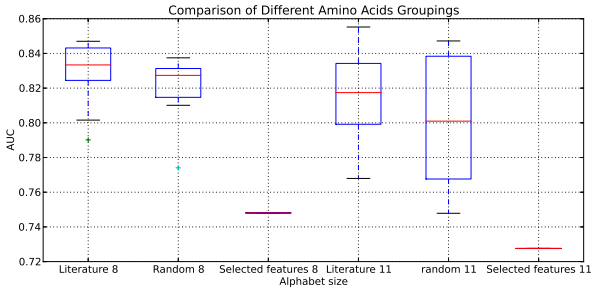


Fig. 2. Comparison of protein-protein interactions with different amino acid alphabet reduction schemes, showing the trade-off between dimensionality and resolution of representation.



reduced amino acid alphabets obtained by random clustering against the alphabet reduction schemes available in the literature based on amono acid properties. Fig. 3Comparison of protein-protein interaction prediction performance of randomly set reduced alphabets and literature derived alphabet reduction schemes at alphabet sizes of eight and eleven. There is significant performance drop when the alphabets are reduced by random clustering. Boxplots are shown across 9 trials in each case. Also shown are performances achieved by feature selection by using a subset of features from the 20 AA representation. There is no equivalent uncertainty measure to be displayedfigure.3 shows that reduced alphabet sizes of eight and eleven, literature derived alphabet reduction schemes achieve significantly higher prediction performances on the protein-protein interaction problem considered. This first check establishes that clustering amino acids based on their chemical and functional properties has merit in representing properties of the resulting proteins.

Fig. 3. Comparison of protein-protein interaction prediction performance of randomly set reduced alphabets and literature derived alphabet reduction schemes at alphabet sizes of eight and eleven. There is significant performance drop when the alphabets are reduced by random clustering. Boxplots are shown across 9 trials in each case. Also shown are performances achieved by feature selection by using a subset of features from the 20 AA representation. There is no equivalent uncertainty measure to be displayed.



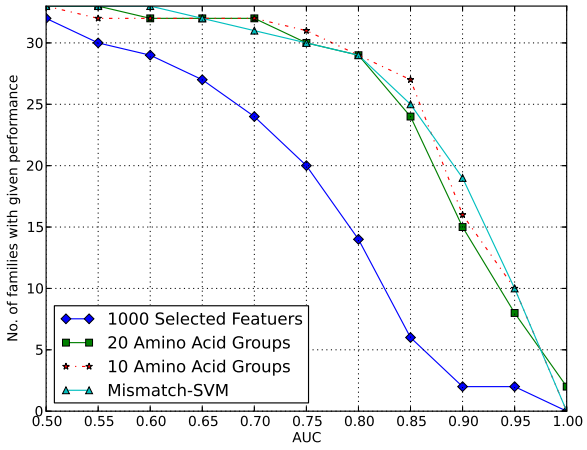
B. Remote Homology Detection

Fig. 4Homology detection with different amino acid alphabet representation schemes. The figure plots the number of SCOP families yielding performances above a certain AUC threshold (x-axis). The Mismatch-SVM results are taken as state-of-the-art performance from [25]figure.4 shows the prediction performance of different amino acid alphabets on the remote homology detection problem. Here we plot the number of families for which the performance exceeded a given AUC threshold, as a function of this threshold. Thus curves towards the top right corner of the graph represent high performance from the system. We note that several reduced alphabet representations achieve performance similar to the published results taken from Leslie *et al.* [25], using a mismatch kernel SVM. Sometimes, the performance is marginally better than the state-of-the-art. Our own implementation used a simple RBF kernel, not particularly tuned to the homology detection problem. Still, with reduced dimensionality, comparable performance can be obtained.

C. Secondary Structure Prediction

The performances on the helix versus non-helix secondary structure classification problem, are shown in Fig. 5Area under ROC curves for a simple secondary structure prediction task of

Fig. 4. Homology detection with different amino acid alphabet representation schemes. The figure plots the number of SCOP families yielding performances above a certain AUC threshold (x-axis). The Mismatch-SVM results are taken as state-of-the-art performance from [25].



classifying α helix structures. Alphabet reduction shows marginal improvement of classification performance on this problem. At the window size used (13 amino acids), the dimensionality of the space is substantially lower than for the other two problems considered. Note, however, the degradation in performance is not sharp, still illustrating the representation-dimensioanlity trade-offfigure.5. This is a simple classifier, working on a window 13 amino acids long. As such, we would not expect good performance in secondary structure prediction. At full representation, an AUC close to 0.775 is obtained, showing the classifier is extracting useful information for prediction. But when the alphabet size is reduced, performance degrades gracefully, rather than suddenly. We note the starting dimension in this problem, $13 \times 20 = 2600$, determined by the chosen window size, is smaller than the interaction and homology problems. Cutting down the alphabet size does not significantly reduce the dimensionality to enhance prediction performance.

IV. Conclusion

Using three different problems of statistical inference applied to protein sequences, we illustrate the compromise that can be struck between fidelity of representation and the effect of dimensionality on the performance of inference algorithms. This compromise is a natural consequence of mapping variable length sequences onto fixed high dimensional spaces. While awareness of it in the form of the curse of dimensionality exists in the literature, most authors stay with an arbitrarily chosen fixed dimension, usually in the full representation of 20 amino acids, for their particular inference problem.

We believe there are interesting implications of our results. While we have considered three popular inference problems, there is a wide range of sequence-based inference problems researchers address. These include motif finding [32], predicting hot-spots [33] along the length of a protein, predicting gene expressions from sequence [34] and so on. Some of these problems are characterized by a substantially worse dimensionality to data ratios than the problems we have considered. As such, the effects of the curse of dimensionality is likely to be higher on these problems, and should be taken into account.

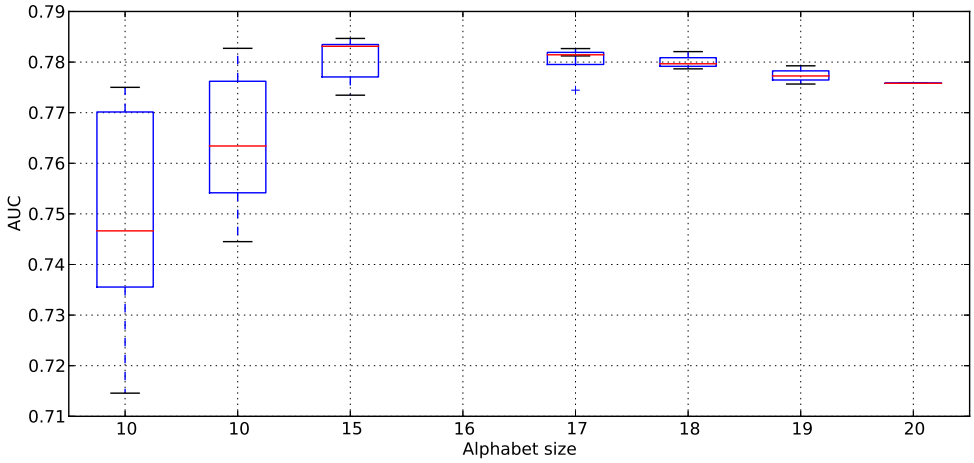
We note that the feature selection based dimensionality reduction by filtering did not produce competitive results. Indeed, on the protein interaction prediction task, feature selection performed substantially worse than any of the alphabet reduction schemes, including the random reduction schemes. Part of the reason is that the feature selection scheme we employed was a simple one (filtering). More powerful ones could be explored as in Scott*et al.*'s work [35]. One could also speculate that when data reduction schemes are employed, the closer one works with raw data, the better performance one gets in terms of inference performance.

Alphabet reduction is but one technique for reducing the dimensionality of a representation. Advanced methods of feature selection [35], [36], principal component projections [37] and sparsity inducing regularizers (*e.g.* L_1 norm regularizer (also known as LASSO), [38]–[40]) have been explored in the machine learning literature and in the context of genomics. While such techniques are primarily statistically motivated, amino acid alphabet reduction offers a biologically meaningful way to deal with the dimensionality issue. Our current work focuses on jointly optimizing alphabet reduction with sparsity inducing regularizers.

References

[1] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, pp. 865 – 884, 1988.
[2] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Journal of Molecular Biology*, vol. 308, no. 2, pp. 397 – 407, 2001.
[3] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Secondary structure prediction with support vector machines," *Bioinformatics*, vol. 19, no. 13, pp. 1650 – 1655, 2003.
[4] C. Cole, J. D. Barber, and G. J. Barton, "The Jpred 3 secondary structure prediction server," *Nucleic Acids Research*, vol. 36, no. suppl 2, pp. W197–W201, 2008.
[5] D. Wieser and M. Niranjan, "Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores," *In Silico Biology*, vol. 9, no. 3, pp. 89–103, 2009.
[6] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting proteinprotein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.
[7] R. E. Bellman, *Dynamic programming*. Princeton University Press, 1957.
[8] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1952.
[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[10] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.
[11] D. Lovell, C. Dance, M. Niranjan, R. Prager, K. Dalton, and R. Derom, "Feature selection using expected attainable discrimination," *Pattern Recognition Letters*, vol. 19, no. 5, pp. 393–401, 1998.

Fig. 5. Area under ROC curves for a simple secondary structure prediction task of classifying α helix structures. Alphabet reduction shows marginal improvement of classification performance on this problem. At the window size used (13 amino acids), the dimensionality of the space is substantially lower than for the other two problems considered. Note, however, the degradation in performance is not sharp, still illustrating the representation-dimensioanlity trade-off.



[12] Y. Saeys, I. Inza, and P. Larraaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[13] E. Peterson, J. Kondev, J. Theriot, and R. Phillips, "Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment," vol. 25, no. 11, pp. 1356–1362, 2009.

[14] S. D. Benson, J. K. Bamford, D. H. Bamford, and R. M. Burnett, "Does common architecture reveal a viral lineage spanning all three domains of life?" *Molecular Cell*, vol. 16, no. 5, pp. 673 – 685, 2004.

[15] M. Munson, L. Regan, R. O'Brien, and J. M. Sturtevant, "Redesigning the hydrophobic core of a four-helix-bundle protein," *Protein Science*, vol. 3, no. 11, pp. 2015–2022, 1994.

[16] C. E. Schafmeister, S. L. LaPorte, L. J. Miercke, and R. M. Stroud, "A designed four helix bundle protein with native-like structure," *Nature Structural & Molecular Biology*, vol. 4, no. 12, pp. 1039–1046, 1997.

[17] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, D. Baker *et al.*, "Functional rapidly folding proteins from simplified amino acid sequences," *Nature Structural Biology*, vol. 4, no. 10, pp. 805–809, 1997.

[18] P. D. Thomas and K. A. Dill, "An iterative method for extracting energy-like quantities from protein structures," *Proceedings of the National Academy of Sciences*, vol. 93, no. 21, pp. 11 628–11 633, 1996.

[19] A. Solis and S. Rackovsky, "Optimized representations and maximal information in proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 38, no. 2, pp. 149–164, 2000.

[20] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, vol. 233, no. 1, pp. 123–138, 1993.

[21] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[22] A. Solis and S. Rackovsky, "Optimized representation and maximal information in proteins," *Proteins*, vol. 38, no. 2, pp. 149–64, 2000.

[23] C. Andersen and S. Brunak, "Representation of protein-sequence information by amino acid subalphabets," *AI Magazine*, vol. 25, no. 1, p. 97, 2004.

[24] J. Yu, M. Guo, C. Needham, Y. Huang, L. Cai, and D. Westhead, "Simple sequence-based kernels do not predict proteinprotein interactions," *Bioinformatics*, vol. 26, no. 20, pp. 2610–2614, 2010.

[25] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, no. 4, pp. 467–476, 2004.

[26] T. Jaakkola, M. Diekhans, and D. Haussler, "Using the fisher kernel method to detect remote protein homologies," in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, 1999, pp. 149–158.

[27] J. A. Cuff and G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," *Proteins: Structure, Function and Bioinformatics*, vol. 34, no. 4, pp. 508–519, 1999.

[28] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146 – 162, 1954.

[29] C. C. Chang and C. J. Lin, "Libsvm : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1 – 27:27, 2011.

[30] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[31] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2006, vol. 207.

[32] T. L. Bailey, "Dreme: motif discovery in transcription factor chip-seq data," *Bioinformatics*, vol. 27, no. 12, pp. 1653–1659, 2011.

[33] N. J. Burgoyne and R. M. Jackson, "Predicting protein interaction sites: binding hot-spots in proteinprotein and proteinligand interfaces," *Bioinformatics*, vol. 22, no. 11, pp. 1335–1342, 2006.

[34] M. A. Beer and S. Tavazoie, "Predicting gene expression from sequence," *Cell*, vol. 117, no. 2, pp. 185 – 198, 2004.

[35] M. Scott, M. Niranjan, D. Melvin, and R. Prager, "Parcel: Feature subset selection in variable cost domains," Cambridge University Engineering Department, Tech. Rep. `CUED/F-INFENG-TR 320`, 1998.

[36] A. Ben-Hur and D. Brutlag, "Sequence motifs: highly predictive features of protein function," *Feature Extraction*, pp. 625–645, 2006.

[37] F. Model, T. Knig, C. Piepenbrock, and P. Adorjn, "Statistical process control for large scale microarray experiments," *Bioinformatics*, vol. 18, no. suppl 1, pp. S155–S163, 2002.

[38] M. Rabinowitz, L. Myers, M. Banjevic, A. Chan, J. Sweetkind-Singer, J. Haberer, K. McCann, and R. Wolkowicz, "Accurate prediction of

hiv-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization," *Bioinformatics*, vol. 22, no. 5, pp. 541–549, 2006.

[39] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[40] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.