1. FINAL PUBLISHABLE SUMMARY REPORT

The two principal tasks of computational analysis of proteins based on their amino acid sequences are the determination of proteins related to one another either by function or by evolutionary development, and the identification of specific amino acid combinations, or motifs, that determine the protein function. Both of these two tasks have largely been addressed by matching amino acid sequences across different proteins. Sequence alignment algorithms compute similarity scores between the amino acid sequences of different proteins that are then used to identify protein subgroups with high within-group similarity. Amino acid combinations observed consistently among proteins of a specific functional subgroup constitute the sequence motifs of the related subgroup, the presence of which indicates membership of the protein in that functional group. The primary challenge in the computational analysis of amino acid sequences is the combinatorial complexity inherent in representing amino acid sequences as words composed of letters from an alphabet of twenty, with each letter corresponding to a different amino acid.

Faced with the daunting prospect of evaluating potentially millions of possible amino acid combinations for functional specificity, we introduce a numerical alternative that characterizes protein structure from amino acid sequences via numerical means using techniques from multi-scale signal decomposition and statistical learning. The proposed framework is based on a notion of hierarchical motif vectors that capture the numerical variation of the local physico-chemical composition along a protein's amino acid sequence. This allows using an extensive library of vector space data processing methods for rigorously computing the similarity of the corresponding amino acid sequence motifs, both in the alignment of amino acid sequences as well as the identification of motifs specific to functional protein groups.

This project starts with developing global and local alignment methods for sequences of motif vectors to establish correspondence between the underlying amino acid sequences. Next, it identifies hierarchical motif vectors that possess functional or structural specificity in select protein groups via quasi-supervised statistical learning. Finally, it formulates a protein function recogniton strategy based on group-specific hierarchical motif vectors.

The experimental results on local as well as global motif vector alignment indicate that the motif vectors adequately characterize the physico-chemical composition along amino acid sequences and allow associating segments sharing similar amino acid configurations at short, mid and long range neighborhoods along their respective sequences. This allows establishing associations between amino acid sequence segments that share similar functions due to amino acid configurations that share similar their physico-chemical properties.

Results on the prediction of N-glycosylation at consensus sequence sites also confirm that the hierarchical motif vectors accurately characterize the physico-chemical configurations at and around amino acid sites for functional significance. Furthermore, the quasi-supervised learning strategy can sort through the prospective sites of activity and identify the ones with real functional potential based on their respective motif vectors. The quasi-supervised learning strategy is especially fitting to biomedical information processing tasks where a relatively small collection of instances are available with experimentally verified attributes against the backdrop of a very large number of unknown prospects. The quasi-supervised learning algorithm successfully separates the probable prospects from the unlikely ones automatically with no user intervention.

A major challenge in generalizing this functional prediction paradigm over all amino acid sites along sequences of a large number of proteins is the computational expense associated with applying the quasi-supervised learning algorithm over large motif vector datasets. While this learning method possesses favorable computational complexity characteristics that allow forming statistically viable predictions over several tens of thousands of vectors simultaneously, the resource requirements become rapidly prohibitive for larger dataset sizes.

As part of this project, an efficient quasi-supervised learning algorithm was developed to address this issue. This algorithm exploited an alternative formulation of the posterior probability estimation scheme that forms the basis of quasi-supervised learning framework, and allowed computing these probabilities using the proximity structure of the individual motif vectors to motif vector clusters. This, in turn, achieved dramatic reductions in the memory requirements to carry out the quasi-supervised learning scheme with comparable prediction accuracy to the original algorithm.

As the first test bed for functional prediction on amino acid sites using the efficient large-scale quasi-supervised learning algorithm on the motif vector data, we have addressed the prediction of the DNA binding sites in human proteins. To this end, we have collected the motif vectors of all amino acid sites of over 20 000 human proteins and carried out large-scale quasi-supervised learning over the resulting 11 208 183 motif vectors to distinguish the ones that correspond to annotated DNA-binding sites from the others. The results indicate that the motif vectors associated with DNA-binding regions were accurately identified along with a large number of sites predicted to be novel DNA-binding sites.

Next, we have used the functional prediction framework developed by this project to identify the amino acid sites along sequences of human proteins possessing annotated DNA-binding sites that exhibited high specificity to that protein group. Since the sites specific to a functional protein group are also the characteristic sites for proteins that belong to that group, locating their motif vectors in the motif vector sequences of novel proteins constitute a strong indication for their inclusion in the same group. Using the specificity measure provided by the framework, we derive a prediction rule to identify the human proteins that were most likely to possess DNA-binding sites among those proteins lacking such annotations. The constructed prediction rule was very accurate at isolating the 654 proteins known to possess DNA-binding sites from the others.

With the successful construction of the prediction scheme to identify unknown members of a given functional protein group described above, all the work related to the project has reached completion as planned. The main outcome of the project is therefore a computational paradigm that represents local amino acid configurations across protein sequences via hierarchical motif vectors, contrasts the motif vector data of proteins belonging to a given functional group to the rest and identifies the amino acid configurations that are characteristic to the group, and predicts new members to the select protein group by evaluating the specificity of their local amino acid configurations to the group through their respective motif vectors.

The ability to accurately predict the likely functional attributes of proteins based on their amino acid sequences offers the potential of targeted protein design whereby sequences for prospective proteins can be evaluated for desired functions *in silico*, and modified systematically to improve the odds. With the methods developed through this project, such a computational paradigm, though at its preliminary stages, appears feasible, and bears wide-ranging potential impact from new generation biomarkers to multi-function molecular agents to be used towards various biological and clinical ends.