A Comparison of Sequence Alignment Algorithms for Measuring Secondary Structure Similarity

L. Gwenn Volkert, *IEEE member* and Deborah A. Stoffer

***Abstract*—Methods for protein secondary structure prediction have improved significantly in recent years. This has lead to enhanced protein homology modeling efforts. Protein homology modeling involves the sub-task of identifying a set of homologous proteins from a protein database when given as input the amino acid sequence of a query protein, with the ultimate goal of using the resulting set of homologous proteins as a starting point for predicting the 3D structure of the query protein. Previous work has indicated that improvements can be made when combining secondary structure sequence alignment using a 3-state structure symbol alphabet together with primary amino acid sequence alignment methods. These approaches typically use a local alignment algorithm. We compare the performance of several dynamic programming alignment algorithms on the task of aligning secondary structure sequences using an 8-state secondary structure alphabet. Our results indicate that the typical use of a local alignment algorithm may not be best when aligning protein secondary structure information.**

Index Terms— Dynamic programming, Pattern classification, Protein secondary structure, Protein function prediction.

I. Introduction

THE determination of protein structure and/or function from primary sequence information remains challenging and costly. As an alternative, researchers continue to develop knowledge-based computational methods such as homology modeling and threading for inference of protein structure and function based on similarities between newly sequenced proteins and databases of existing protein information, for a review of various approaches see [1]. Many of the more recent knowledge based approaches to the structure prediction problem use a measure of secondary structure similarity as a part of the overall process [2-7]. The measurement of secondary structure similarity between two or more proteins can be made in many ways. In the late 1990's McGuffin et al. compared protein domain secondary structure similarity scoring methods for eleven different pairwise similarity approaches [8]. In general it was shown that secondary structure similarity was not a sufficient method when taken in isolation for solving the protein homology-modeling problem. More recently, several attempts have been made that combined secondary structure similarity with primary sequence based homology methods with modest success [9, 3, 10]. The secondary structure similarity measure employed in these particular approaches is determined by aligning sequences of secondary structure state symbols that represent the secondary structures of the proteins being compared. This form of the alignment problem is typically simplified by reducing the variety of secondary structures represented from the 8-state standard developed for the DSSP program [11] to a simpler three-state representation. The 8-states defined by the DSSP method account for three different types of helices, two different types of strand related structures, 2 turn structures and a catch-all state for residues in the loop/coil regions, while the 3-state alphabet, are often referred to as α, β, and γ (or coil), reduces the representation to just residues involved in helices, strands, and other structures (e.g.,loops/coils or unknown) respectively.

Secondary structure sequence alignment is typically implemented as a Smith-Waterman *local* alignment algorithm [12]. Preliminary work using a Needleman-Wunsch *global* alignment algorithm [13] as part of our own combined protein homology approach, led to the observation that a closer examination of the performance of different alignment algorithms was warranted when working with sequences representing secondary structure. This paper describes our initial efforts in this area and highlights the need for a more thorough comparison of alignment algorithms for secondary structure sequences. We compare six variations of dynamic programming (DP) based alignment algorithms for aligning secondary structure sequence data using a database of known protein secondary structure sequences. Our alignment algorithm variations include local, global and semi-global approaches each implemented to support either linear or affine gap penalty functions. The results indicate that the typical choice of a Smith-Waterman local alignment algorithm may not be the best choice for aligning secondary structure

Manuscript received June 18, 2004.

L. Gwenn Volkert is with the Department of Computer Science at Kent State University, Kent, OH 44240 USA phone: 330-672-9037; fax: 330-672-7824; (e-mail: volkert@ cs.kent.edu).

Deborah Stoffer is with the Department of Computer Science at Kent State University, Kent, OH 44240 USA (e-mail: dstoffer@cs.kent.edu).

0-7803-8728-7/04/$20.00 ©2004 IEEE

sequences. We conducted the comparisons using a dataset of 210 globular proteins representing 113 different SCOP superfamilies [14] and 27 different CATH homologous superfamilies [15]. The test dataset is derived from a globular protein dataset originally developed by Michie et al. [16] for use with their automated class assignment protocol. The same dataset was also used in other efforts at automating structural class prediction from sequence data [17]. The dataset was comprised of a set of 210 protein domains classified into three classes based on their secondary structure motifs. Throughout the rest of this paper we will use the term protein to mean a protein domain, with the understanding that the Protein Data Bank (PDB) IDs used will always indicate the specific domain we are using.

The three protein classes represented in the test dataset are referred to as mainly-α, mainly-β, and mixed-α/β, that contain 59, 79, and 64 proteins respectively, The secondary structure sequences of all 210 proteins in this dataset were derived using the DSSP method [11]. For three sets of query proteins containing five proteins each we ran each of six alignment algorithm variations to generate alignment scores against all the proteins in our test As with all dynamic programming algorithms for sequence alignment, the scoring mechanism is dependent on the scoring matrix. We use an 8-state scoring matrix designed by modifying the 3-state scoring matrix used in the secondary alignment algorithm implemented by Wallqvist et al. [3].

II. BACKGROUND

Homology modeling is based on the observation that proteins with similar sequences tend to fold into similar structures. The goal of homology modeling programs is to generate a set of homologous proteins to an unknown "query" protein. The resulting homologous set will then be used to predict the 3D structure by comparing the query protein sequence with the sequences of the homologous proteins for which X-ray or NMR based structure data is known. Homology modeling is supported by the empirically determined fact that at least 66% of the known proteins having a similar structure also have a similar function [18]. Homology modeling starts with a template identification step based either on sequence to sequence alignment, or sequence-to-structure alignment (threading). In the interest of space we restrict the remainder of our discussion to homology modeling based on sequence-to-sequence alignment. Sequence alignment based on pairwise primary sequence identity works well for detecting homologs that have a high degree of sequence similarity but usually does not work for homologs with a low degree of sequence similarity. Primary sequence alignment approaches suffer from an inability to differentiate between protein sequences that differ significantly in primary sequence but never the less code for very similar secondary and or tertiary structures [19]. The homologous sets identified by using primary sequence alignment approaches often miss some, if not many, of the homologous proteins. These difficult to detect homologs are often referred to as those in the *twilight zone* and are typically defined to be homologous proteins with less than 30% primary sequence similarity [19, 20].

Because protein function is known to be related to its 3D (or tertiary) structure, which is constrained by its 2D (or secondary) structure, some protein homology modeling approaches have been enhanced to combine secondary structure similarity measure together with primary sequence alignment results to improve the coverage of the resulting homology set [10, 9]. The problem of secondary (and tertiary) protein structure prediction is inherently difficult and is often approached with computational intelligence methods. These programs typically combine many of the same heuristic algorithms as used in homology modeling. Many structure prediction methods employ artificial neural networks (ANNs) [21-27], and are typically considered to be the most accurate [28, 29]. Other heuristic based approaches that have been investigated include hidden Markov models [30-32], support vector machine approaches [33] and others [34-36].

TABLE I
8-STATE STRUCTURE SYMBOLS DEFINED BY DSSP WITH THREE TYPICAL 3-STATE REDUCTIONS

DSSP SYMBOL	STRUCTURE DESCRIPTION	CASP REDUCTION	SIMPLEST REDUCTION	SSAS* REDUCTION
H	alpha-helix	α	α	α
G	3/10-helix	α	γ	α
I	pi-helix (rare)	γ	γ	γ
E	extended stand	β	β	β
B	beta bridge	β	γ	γ
T	H-bonded turn	γ	γ	γ
S	bend	γ	γ	γ
.	loop/coil or unknown	γ	γ	γ

* The SSAS program is described in [3].

Secondary structure sequence similarity procedures are often built within structural prediction programs. For example, in early work by Russell et al., a fast algorithm was developed for generating all exact matching alignments under the constraint of a maximum number of insertions or deletions. The sequences being aligned were comprised of helix (H) and strand (B) secondary structure state symbols. In this approach each H represented an entire α-helix and each B represented a β-strand in the proteins to be aligned [5]. More recent secondary structure sequence representations maintain the same reside sequence length as the primary sequence but use a secondary structure state alphabet such as those illustrated in Table I. In general secondary structure sequences are represented using either an 8-state alphabet or various 8-3 state reduction alphabets. Each amino acid of a protein sequence is represented in a secondary structure sequence by a symbol from a secondary structure alphabet yielding a secondary structure sequence of the same length as the

original amino acid sequence.

Many combination based approaches for the identification of homologous proteins employ either BLASTp or PSI-BLAST as a first step in the combined process. These programs are generally very effective at identifying homologous proteins that have greater than 30% amino acid similarity. Methods that seek to identify homologous proteins with a low degree of sequence similarity can incorporate secondary structure information into their approaches. These can generally be divided into two approaches, either through the use of secondary structural information as an additional constraint on the primary alignment results [28, 4] or by using secondary structure information as the primary source of information [37, 6]. The interest in using secondary structure information in homology programs continues to grow as improvements in secondary structure prediction programs continue to be made [38, 39]. The results of a secondary structure prediction program can then be used to generate secondary structure similarity scores that can then be combined with primary sequence based homology prediction. There are numerous secondary prediction programs currently available (see [40] for a review), we briefly mention just three of them here as examples of the variety of approaches available. Jpred [41] and SSPro8 [22] are both ANN approaches and SAM-T99 [32] is a Hidden Markov model approach. Jpred uses a two-level neural network algorithm together with PSI-BLAST derived multiple sequence alignment profiles to predict the secondary structure. SSPro8 use bidirectional recurrent neural networks and PSI-BLAST derived profiles. In each case the neural networks are trained using protein sequence multiple alignments and known secondary structures from the PDB. SAM-T99 takes a single sequence and iteratively develops a hidden Markov model (HMM) from the sequence and homologs found using the HMM in a database search. The trained HMM is then used to predict the protein secondary structure for an unknown amino acid sequence.

Clearly secondary structure sequence alignment is not a new concept as it is often an integral part of both homology modeling programs and secondary structure prediction programs. What is surprising is that little justification for the specific type of alignment algorithm (i.e. global or local) is given. In general Smith-Waterman local alignment algorithms are used with a passing reference to their appropriateness for addressing alignment of protein primary sequences. The main difference between global and local alignment is the trade-off between recognition of overall similarity (generally reserved for aligning long stretches of DNA between different genomes) and recognition of evolutionarily conserved regions of DNA or protein, which are relatively short and are often flanked by region that will not easily align. Given the biological and physical differences between amino acid sequences and secondary structure sequences a comparison of advantages and disadvantages of the alignment algorithm variations is warranted.

III. Methods

A. The Algorithms

We have implemented and compared the performance of three different DP based alignment methods each with two different methods of determining gap penalties. The specific DP approaches compared are listed in Table II. By definition a DP alignment algorithm will generate the best alignment of two sequences, S_1 and S_2, according to the measurement criteria given in the scoring matrix. The score of each alignment is calculated as the sum of the scores of each aligned pair of sequence symbols s_1 and s_2, where the pairwise score is given by a value in a scoring matrix M. The scoring matrix supplies a pairwise score for all possible pairs of secondary sequence symbols.

TABLE II
List of Alignment Algorithms

Algorithm	Approach
Global+Affine	Global
Global+Linear	
Local+Affine	Local
Local+Linear	
Semi-Global+Affine	Semi-Global
Semi-Global+Linear	

Starting with the sequence alignment algorithm package written by Rolf Backofen and Sebastian Will, three different dynamic programming based alignment algorithms were implemented to compare secondary structure descriptors of a target set of proteins obtained from the PDB [42]. A secondary structure descriptor is a vector of characters where each character represents the type of secondary structure each amino acid participates in. The algorithms use an 8-state alphabet containing the characters H (residue participates in alpha helix), B (residue in isolated beta-bridge), E (residue is part of an extended strand and participates in beta ladder), G (residue in 3/10 helix), I (residue in a pi helix), T (residue in hydrogen bonded turn), S (residue in bend), or "." (residue is part of a loop/coil region, or unknown) as defined in the Definition of Secondary Structure of Proteins (DSSP) program [11]. The six variations we have tested are listed in Table II.

Each algorithm accepts as input a query protein in secondary structure sequence character string format, a similarity scoring matrix, and the name of a file containing protein secondary structure sequences to align the query protein with. The similarity score for each alignment is written to an output file in rank order based on the alignment similarity score along with the PDB protein ID, and the protein secondary structure of the target.

B. The Scoring Matrix

The scoring matrix used in all of the alignment algorithms compared was extended to support an 8-state alphabet from the 3-state similarity scoring matrix developed for the SSAS project [3]. The original SSAS scoring matrix

was calculated from log-odds scores based on three-dimensional alignments obtained from 3D-ali data bank. The values for the extended states were determined by either copying or adjusting the values from the most similar scoring pair as indicated by the CASP 8-state to 3-state reduction (see Table I). The secondary structure strings of the globular proteins in general are mostly composed of alpha-helices (H) and extended-strands (E), and residues participating in loop/coils. The secondary structure states, G, B, T, and S are less common and I's are rare.

It is likely that recalculating the log-odds scores in terms of an 8-state alphabet would produce a more accurate similarity matrix. We plan to explore this in future work. Favorable pairings are assigned positive scores and negative scores are assigned to structural elements least likely to be found together. Implementation of the gap penalty versions were uniformly implemented for the three types of alignment algorithms such that the gap opening and elongation parameters for the affine gap penalty versions were set to -12 and -2 respectively, and for the linear gap penalty versions a gap penalty of -2 was used.

TABLE III
SECONDARY STRUCTURE SIMILARITY MATRIX

	H	G	I	E	B	T	S	other
H	2	1	1	-15	-4	-4	-4	-4
G	1	3	1	-15	-4	-4	-4	-4
I	1	1	3	-15	-4	-4	-4	-4
E	-15	-15	-15	4	-4	-4	-4	-4
B	-4	-4	-4	-4	2	1	1	1
T	-4	-4	-4	-4	1	2	1	1
S	-4	-4	-4	-4	1	1	2	1
other	-4	-4	-4	-4	1	1	1	2

TABLE IV
PDB CODES OF PROTEINS (DOMAINS) USED ORGANIZED BY STRUCTURAL class with proteins used as queries marked with asterisks.

	1eca_	1mbd_	3sdhA	1hbg_	1thbA	1mba_	1lh1_ *	1ithA
	1cpcA	1cpcB	1colA	1lmb3	1lccA	1r69_	1utg_	1aca_
	1fiaA *	2wrpR	1hddC	1rro_	1osa_	4icb_	2sas_	351c_
mainly-α	3c2c_	2mtaC	1cc5_	1c5a_	1prcC *	256bA	2ccyA	2hmqA
	1lpe_	1bbhA	1fha_	1ropA	1lfb_	1aep_	2tmvP *	1rcb_
	3inkC	1prcL	1prcM	1ppa_	1gluA	1hyp_	1lis_ *	1ltsC
	1huw_	1rfbA	1d66A	1ysaC	1acp_	1bha_	1bgc_	1ribA
	1cmbA	1poc_	1ltsA					
	1ifc_	1mdc_	1opaA *	1stp_	1aveA	1bbpA *	1hbq_	1mup_
	2por_	1omf_	2sga_	2pkaA	2rhe_	1cd8_	1cid_	1tlk_
	1fc2D	1noa_ *	1ttaA	1ttf_	1cdb_	1ten_	1cobA	1plc_
mainly-β	1paz_	1aaj_	1aizA	1hoe_	2stv_	1tnfA	1tme1	2plv1
	1bbt2	2plv2	1bbt3	2plv3	4sbvA	2tbvA	1bmv1 *	2ctvA
	2ltnA	2ayh_	1sltA	2rspA	1hivA	1gpr_	1nscA	2bat_
	2sim_	4fgf_	1i1b_	1tie_ *	3ebx_	1cdtA	1fas_	1atx_
	1bds_	1egf_	2tgf_	2tgi_	1pdgA	1hcc_	1tpm_	1bgh_
	4sgb1	2ltnB	4htc1	1lyaA	2bpa2	1lab_	3monA	1tfi_
	1cauB	1shg_	1pnj_	2cpl_	1bw3_	1ltsD	2sns_	
	1xis_	5timA	1nar_	2mnr_	1chrA	1fbaA	1gox_	5p21_
	3chy_	1etu_	1ofv_	4fxn_	1cseE	1cde_ *	2trxA	3trx_
	1gp1A	4dfrA	1ak3A	3adk_	2ctc_	2cmd_	1ipd_	7icd_
	1nipA	1tml_	1gps_	2ovo_	1tgs1	1gatA *	1ptf_	2bopA
mixed α, β	1ctf_	1fxd_	2nckL	3rubS	1pba_	1aps_	1cse1	2sic1
	2rn2_ *	1aak_	7rsa_	1onc_	1fus_	1brnL	1pgx_	1frrA *
	1ubq_	3monB	1vil_	3il8_	1fkb_	2msbA	5pti_	1adn_
	1zaaC	1shaA	2pna_	3cla_	1eaf_	1cewI	2tscA	1rveA
	1stf1	3b5c_	1pkp_	5fd1_	2dnjA *	1mat_	1pyaB	1hgeB

C. The Test Dataset

We conducted 15 runs of each algorithm using a dataset of 210 globular proteins (see Table IV). The dataset consists of proteins originally classified as mainly-α, mainly-β, α/β, or α+β by Michie et al. [16]. Of these 210 proteins, 56 were originally classified as mainly-α, 75 as mainly-β, 26 as α/β, and 53 as α+β. For our experiments the mainly-α class contains 59 proteins, the mainly-β class contains 79 proteins, and the α/β and α+β classes were combined into one mixed-αβ class of 72 proteins after reassigning several proteins from this class to the mainly-α mainly-β class based on current SCOP and CATH information. Five structures from each of the three classes were randomly chosen to be used as query proteins for each experiment. For each protein in the test dataset we also obtained the most recent SCOP super-family and CATH homology codes for use in our analysis. Five of the query proteins have homologous proteins in the test dataset as determined by their CATH and SCOP superfamily classifications. The five query proteins along with their respective CATH and SCOP superfamily codes are given in Table V.

TABLE V
QUERY PROTEINS AND CATH AND SCOP CODES

Class	PDB ID	Sequence Length	CATH superfamily	SCOP superfamily
mainly-α (59)*	1lh1_	130	10	46458
	1fia_A	79	60	48283
	1prc_C	138	10	48695
	2tmv_P	154	70	47195
	1lis_	131	10	47082
mainly-β (79)*	1opa_A	133	20	50814
	1bbp_A	173	20	50814
	1noa_	113	230	49319
	1bmv1	123	20	88633
	1tie_	166	50	50386
Mixed-αβ (72)*	1cde_	209	170	53328
	1gat_A	60	10	57716
	2rn2_	155	10	53098
	1frr_A	95	30	54292
	2dnj_A	253	10	56219

* the number in parenthesis indicates the total number of proteins of the indicated class in the test dataset

D. Experiment Description

For each of the fifteen query proteins, five from each of the three different globular classes, alignments were obtained for all of the 210 secondary structures in the test dataset described above. The resulting alignments were ranked according to their alignment score. For each query protein the number of proteins in each of the globular classes (mainly-α, mainly-β, and mixed-αβ) that ranked within the bounds of the number of proteins in the globular class of the query protein was recorded along with the similarity score generated by the algorithm. As noted in Table III there are 56 mainly-α proteins, 75 mainly-β proteins, and 79 mixed-αβ proteins. We also recorded for each query protein the highest and lowest ranking alignment achieved for a protein of each class.

IV. RESULTS

We have tested six different alignment algorithms to compare their effectiveness for detecting homologous relationships between proteins based only on secondary structure descriptors (i.e. vectors of secondary structure states). The algorithms were applied to our test dataset of 210 globular proteins where the structures are known and the corresponding sequences have been previously categorized into particular protein families. We used as our query sequences 15 proteins randomly selected from our test set such that there were five proteins selected from each of the three classes. For each query protein submitted the algorithm returns a list of the target proteins ranked by their similarity score.

For each alignment algorithm we calculate the percentage of proteins, of the same class as the query, that rank in the top *n* positions, where *n* is the number of proteins in the class that the query protein is a member of. We also record the percentage of proteins of the other two classes that rank in the top *n* positions. We label each of the possible relationships as *i/j*, where *i* represents the class of proteins being counted within the top positions possible for the class that the query protein is a member of, where *j* identifies this expected class. For example, if the ranked list of aligned proteins for an alpha query included 51 of the 56 possible alpha proteins along with 5 ab-mixed proteins in the top 56 positions of the ranked list the α/α score would be 91%, the β/α score would be 0% and the mixed-αβ/α score would be 9%. This recording method generates three values for each of the 15 query proteins tested giving a total of 45 values for each algorithm. We condense this information into a chart depicting the average percentage for the three sets of five query proteins representing each of the three classes, mainly-α, mainly-β, or mixed-αβ.

In Fig. 1 each grouping of bars gives the average percentage of same class rankings for each of the three sets of query proteins mainly-α, mainly-β, and mixed-αβ respectively. For example, the first grouping (labeled α/α) illustrates the average percentage of mainly-α proteins that were ranked above the highest ranked protein from either of the other classes, for each of the six algorithms compared. The first observation to note is that mainly-α and mixed-αβ proteins achieve overall higher similarity scores when aligned using the SEMI-GLOBAL+AFFINE algorithm and the lowest scores for either the LOCAL+LINEAR or the LOCAL+AFFINE algorithm. Our intuition is that since global alignment methods are intended to maximize the alignment over the length of the entire sequence more of the individual helices end up aligned with

each other since the individual helices are arranged sequentially along the sequence. This advantage would not be as easily exploited for the local style alignment algorithm, regardless of gap penalty function. In the case of mainly-β proteins it is plausible that the local alignment algorithm is able to achieve similarity scores higher than those achieved using a global style algorithm since the residues participating in beta strands will not necessarily be side-by-side. The semi-global alignment algorithm is able to ignore leading and trailing gaps. This feature in combination with the observations noted above may be the reason for the overall good performance of the semi-global style.

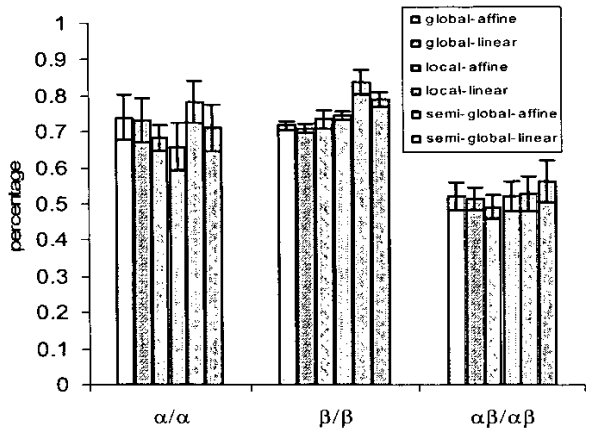
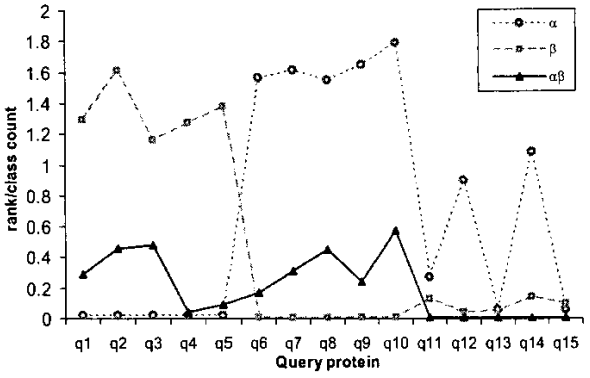


Figure 1. The chart shows the average percentage including standard deviation of the proteins of the same classification as the query proteins for the six different alignment algorithms compared. In general the best performance is observed when the SEMI-GLOBAL alignment algorithm is used.

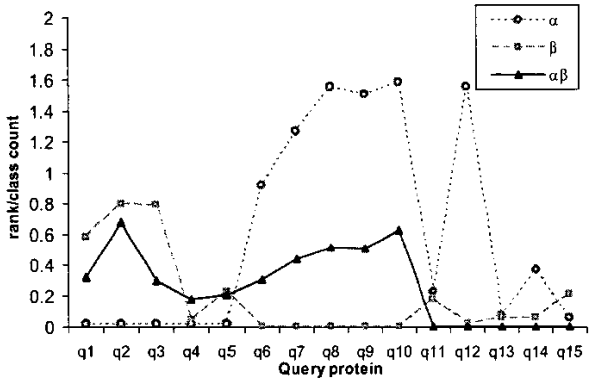
To obtain a measure of how likely it is for proteins from different classes to align better than proteins of the same class, the rank of the highest ranked alignment to a protein of a different class than the query protein is divided by the count of the number of proteins of the query class. In Figure 2 these results are illustrated in charts (a), (b), and (c), representing the data from the GLOBAL+AFFINE, SEMI-GLOBAL+AFFINE, and LOCAL+AFFINE runs respectively. A lower rank/count measure indicates a higher ranking. It is not surprising that for all alignment algorithms and for each class of proteins the highest ranked protein is of the same class, as it is expected that the query protein should result in the best alignment when aligned to itself. In some cases, proteins of a different class achieve relatively high rankings. This is most apparent with the mixed-αβ class query proteins since all of the points along the mixed-αβ line are less have a value of less than 1.0. We observed that the rank/count measures are consistently lower (worse) for the mixed-αβ values for all mainly-α queries and are lower for three of the five mainly-β queries measured using the LOCAL+AFFINE algorithm. It should be noted that the highest ranked mixed-αβ protein for the q9 query has a CATH classification of mainly-β and a SCOP classification of mixed-αβ.

A comparison of the β line in the three charts indicates that the GLOBAL+AFFINE algorithm is less likely than the other algorithms to rank an alignment with a mainly-β protein high when the query protein is a mainly-α protein.

(a) Global+Affine



(b) Semi-Global+Affine



(c) Local+Afffine

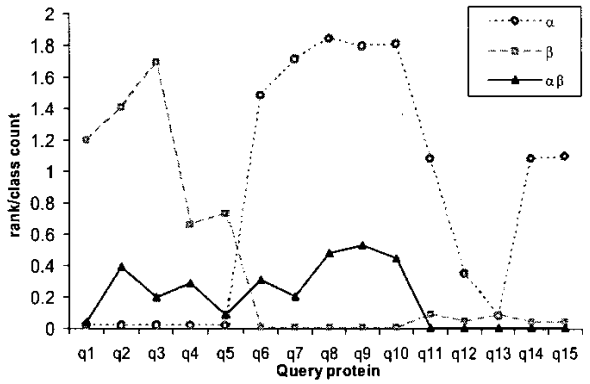


Figure 2. For each of the 15 query proteins (q1-q15), the highest ranked alignment to a protein of a different class than each of the query proteins divided by the count of the number of proteins in the query class is shown. This measure indicates differences in how the alignment algorithms perform and indicates that the use of a LOCAL+AFFINE alignment approach is not necessarily the best choice.

Five of the query proteins have CATH and SCOP superfamily codes represented multiple times in the dataset. These homologs are shown in Table VI along with the alignment ranking obtained using the SEMI-GLOBAL+AFFINE alignment algorithm. For each of the mainly-α and mainly-β query proteins the highest ranked alignment, (identified in the table with an ALIGNMENT RANK value of 2) is achieved with a protein identified as a CATH and SCOP homolog. The ALIGNMENT RANK is 2 since in all the rank 1 alignment are to the query protein itself. The rankings for the query proteins that have SCOP homologs in the test dataset are given in Table VI.

TABLE VI
RANK OF HOMOLOGOUS PROTEINS FROM SEMI-GLOBAL+AFFINE ALIGNMENT

QUERY PROTEIN PDB ID	SCOP BASED HOMOLOG PDB ID	ALIGNMENT RANK
	1mbd:_	2
	1hbg:_	4
1lh1_	1mba:_	6
(mainly-α)	1ith:A	7
	1eca:_	8
	3sdh:A	10
	1cpc:A	11
	1lis_	12
	1bmv:1	2
	2plv:3	7
1bmv:1	1tme:1	10
(mainly-β)	1bbt:2	15
	2tbv:A	18
	4sbv:A	45
	2plv:1	62
	2plv:2	63
	1hbq:_	2
1bbp:A	1mup:_	3
(mainly-β)	1mdc:_	13
	1ifc:_	16
	1opa:A	17
	1mdc:_	2
1bbp:A	1opa:A	3
(mainly-β)	1ifc:_	19
	1hbq:_	24
	1mup:_	49
1gat:A	1glu:A	10
(mixed-αβ)		

V. CONCLUSION

Many structural prediction and homology modeling methods use various forms of ANNs in their implementations. This powerful computational intelligence approach often is combined with primary and secondary structure alignment algorithms to improve the quality of the results. The typical approach used for alignment of secondary structure sequences for proteins is a Waterman style local alignment algorithm using a 3-state secondary structure alphabet. Our comparison of six different alignment algorithms for aligning protein secondary structure sequences, indicates that the accuracy of the structural prediction and homology modeling methods is likely to be affected by the type of alignment algorithm used. While some research has focused on comparing different types of computational intelligence methods for structural prediction and homology modeling, our results indicate that the type of secondary structure sequence alignment algorithm used also deserves careful scrutiny.

We have compared local, global, and semi-global approaches with linear and affine gap penalty functions using the DSSP 8-state secondary structure alphabet. The preliminary results indicate that semi-global alignment algorithms may be a better choice when aligning secondary structure sequences. The preliminary nature of this comparison leaves many avenues for future work. To date we have experimented only with globular proteins, using a secondary structure similarity matrix for which only 3 of the 8 possible state symbol pairings are based on derived probabilities. Thus the similarity matrix used here may not be the best choice. Additionally, it is not likely that the best scoring matrix for globular proteins will be appropriate for other types of proteins. Future work will need to address possible differences in how secondary structure sequence alignment works on fibrous and membrane type proteins as well.

Overall, the differences exhibited throughout the limited set of runs reported indicate that a more thorough comparison of secondary structure sequence alignment is warranted. Never the less the limited results presented enable the framing of additional questions about the usefulness of secondary structure sequence alignment. The datasets used in this report are not the ideal choice for testing protein homology modeling in the twilight zone (those exhibiting <30% amino acid sequence homology). We are in the process of running a set of more detailed experiments with datasets that specifically address the difficulties encountered with matching twilight-zone proteins. A more detailed studied is clearly warranted and will aid researchers in making intelligent choices when incorporating secondary structure alignment into a variety of bioinformatics applications.

ACKNOWLEDGEMNTS

We thank the conference reviewers for their time and effort and are grateful for their useful comments and suggestions.

REFERENCES

[1] S. C. Teichman, C. Chothia, and G. M. Church, "Advances in structural genomics.," *Current Opinions in Structural Biology*, vol. 9, pp. 390-399, 1999.

[2] H. Xu, R. Aurora, G. D. Rose, and R. H. White, "Identifying two ancient enzymes in Archaea using predicted secondary structure alignment," *Nat Struct Biol*, vol. 6, pp. 750-4., 1999.

[3] A. Wallqvist, Y. Fukunishi, L. R. Murphy, A. Fadel, and R. M. Levy, "Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases," *Bioinformatics*, vol. 16, pp. 988-1002, 2000.

[4] B. Rost, R. Schneider, and C. Sander, "Protein fold recognition by prediction-based threading," *J Mol Biol*, vol. 270, pp. 471-80., 1997.

[5] R. B. Russell, R. R. Copley, and G. J. Barton, "Protein fold recognition from secondary structure assignments," *Proc. 28th Hawaii. Int. Conf. Sys. Sci. IEEE Press*, vol. 5, pp. 302-311, 1995.

[6] R. Aurora and G. D. Rose, "Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparision.," *Proc Natl Acad Sci U S A*, vol. 95, pp. 2818-2823, 1998.

[7] R. Lüthy, A. D. McLachlan, and D. Eisenberg, " Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities.," *Proteins*, vol. 10, pp. 229-239, 1991.

[8] L. J. McGuffin, K. Bryson, and D. T. Jones, "What are the baselines for protein fold recognition?," *Bioinformatics*, vol. 17, pp. 63-72, 2001.

[9] C. Geourjon, C. Combet, C. Blanchet, and G. Deleage, "Identification of related proteins with weak sequence identity using secondary structure information," *Protein Science*, vol. 10, pp. 788-797, 2001.

[10] E. Bindewald, A. Cestaro, J. Hesser, M. Heiler, and S. C. Tosatto, "MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification," *Protein Eng*, vol. 16, pp. 785-9., 2003.

[11] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-637., 1983.

[12] T. F. Smith and M. S. Waterman, "Identification of common molecular sequences," *J Mol Biol*, vol. 147, pp. 195-197, 1981.

[13] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *J Mol Biol*, vol. 48, pp. 443-453, 1970.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures.," *J Mol Biol*, vol. 247, pp. 536-540, 1995.

[15] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH- A Hierarchic Classification of Protein Domain Structures.," *Structure*, vol. 5, pp. 1093-1108, 1997.

[16] A. D. Michie, C. A. Orengo, and J. M. Thornton, "Analysis of domain structural class using an automated class assignment protocol," *J Mol Biol*, vol. 262, pp. 168-85., 1996.

[17] C. T. Zhang and R. Zhang, "A new criterion to classify globular proteins based on their secondary structure contents," *Bioinformatics*, vol. 14, pp. 857-65., 1998.

[18] F. S. Domingues, W. A. Koppensteiner, and M. J. Sippl, "The role of protein structure in genomics," *FEBS Lett*, vol. 476, pp. 98-102., 2000.

[19] B. Rost, "Twilight zone of protein sequence aligments," *Protein Engineering*, vol. 12, pp. 85-94, 1999.

[20] R. F. Doolittle, *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences.* Mill Valley, CA, USA.: University Science Books, 1986.

[21] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J Mol Biol*, vol. 202, pp. 865-84., 1988.

[22] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins*, vol. 47, pp. 228-35., 2002.

[23] L. H. Hung and R. Samudrala, "Accurate and automated classification of protein secondary structure with PsiCSI," *Protein Sci*, vol. 12, pp. 288-95., 2003.

[24] L. H. Holley and M. Karplus, "Neural networks for protein structure prediction," *Methods Enzymol*, vol. 202, pp. 204-24., 1991.

[25] J.-M. Chandonia and M. Karplus, "New methods for accurate prediction of protein secondary structure," *Proteins*, vol. 35, pp. 293-306., 1999.

[26] S. K. Riis and A. Krogh, "Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments," *J Comput Biol*, vol. 3, pp. 163-83., 1996.

[27] B. Rost, "PHD: predicting one-dimensional protein structure by profile-based neural networks," *Methods Enzymol*, vol. 266, pp. 525-39., 1996.

[28] L. Jaroszewski, W. Li, and A. Godzik, "In the search for more accurate alignments in the twilight zone.," *Protein Science*, vol. 11, pp. 1702-1713, 2002.

[29] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins*, vol. 40, pp. 502-11., 2000.

[30] C. Bystroff, V. Thorsson, and D. Baker, "HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins," *J Mol Biol*, vol. 301, pp. 173-90., 2000.

[31] R. Karchin, M. Cline, Y. Mandel-Gutfreund, and K. Karplus, "Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry," *Proteins*, vol. 51, pp. 504-14., 2003.

[32] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov models for detecting remote protein homologies," *Bioinformatics*, vol. 14, pp. 846-56., 1998.

[33] J. Guo, H. Chen, Z. Sun, and Y. Lin, "A novel method for protein secondary structure prediction using dual-layer SVM and profiles," *Proteins*, vol. 54, pp. 738-43., 2004.

[34] K. Pawlowski, L. Rychlewski, B. Zhang, and A. Godzik, "Fold predictions for bacterial genomes," *J Struct Biol*, vol. 134, pp. 219-31., 2001.

[35] A. A. Salamov and V. V. Solovyev, "Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments," *J Mol Biol*, vol. 247, pp. 11-5., 1995.

[36] R. D. King, M. Saqi, R. Sayle, and M. J. Sternberg, "DSC: public domain protein secondary structure predication," *Comput Appl Biosci*, vol. 13, pp. 473-4., 1997.

[37] R. B. Russell, R. R. Copley, and G. J. Barton, "Protein fold recognition by mapping predicted secondary structures," *J Mol Biol*, vol. 259, pp. 349-65., 1996.

[38] T. N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. P. Gippert, and O. Lund, "Prediction of protein secondary structure at 80% accuracy," *Proteins*, vol. 41, pp. 17-20., 2000.

[39] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *J Mol Biol*, vol. 232, pp. 584-99., 1993.

[40] B. Rost, "Review: protein secondary structure prediction continues to rise," J Struct Biol, vol. 134, pp. 204-18., 2001.

[41] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, "JPred: a consensus secondary structure prediction server," Bioinformatics, vol. 14, pp. 892-3., 1998.

[42] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Res, vol. 28, pp. 235-42., 2000.