The Context-Dependence of Amino Acid Properties*

Thomas R. Ioerger
Department of Computer Science
Texas A&M University
College Station, TX 77843
ioerger@cs.tamu.edu

Abstract

One of the current limitations of using sequence alignments to identify proteins with similar structures is that some proteins with similar structures do not have significant sequence similarity by identity. One way to address this "hidden-homology" problem is to match amino acids based on their chemical and physical properties. However, the amino acid properties overlap, creating orthogonal dimensions of similarity, the relative strengths of which are ambiguous. It has been observed that the role an amino acid plays (and hence the property that is important) at a site in a protein depends on its secondary and tertiary environment. To approximate and take advantage of this dependence on context for improving the sensitivity of alignments of proteins whose structures are unknown, we propose a surrogate definition of context based on the pattern of hydropathy in a small window of contiguous neighbors surrounding each amino acid. We present the results of an experiment in which a search-based program iteratively tests and selects various properties in independent contexts, and incrementally increases the ability of sequence alignments to detect relationships among distantly-related proteins. The method is shown to perform better than using the MDM78 substitution table for partial match scores.

Introduction

The structure of a protein is determined not only by its amino acid sequence, but also by the nature of those amino acids. Each of the 20 amino acids has a unique combination of physical and chemical properties based on its side-chain. Examples of physical properties include *volume, length, backbone constraint (ϕ/ψ angles), side-chain flexibility (χ angles)*, and *branching structure*. Examples of chemical properties are *hydrophobicity, polarity, charge, reactivity, solubility, aromaticity*, and the *ability to participate in hydrogen bonds*. Properties such as these determine what role each amino acid plays in stabilizing a protein structure, and what other amino acids can substitute for it.

Some of the earliest evidence for the importance of amino acid properties came from observations of non-uniform substitution patterns. Dayhoff et al. (1972) collected statistics on the rate of replacement between each pair of amino acids in multiple alignments of various protein families. They found that some substitutions occurred more frequently than expected, and others occurred less frequently than expected, relative to the individual frequencies of occurrence. It was observed that the amino acids tended to fall into exchange groups such that, within a group, amino acids exchange with increased frequency, but between groups, exchanges were suppressed. Dayhoff interpreted these exchange groups by identifying properties that were common among all the members of each group: *hydrophobic* (met ile leu val), *positive* (his arg lys), *aromatic* (phe trp tyr), *small* (ala asp glu gly asn pro gln ser thr), and *reactive* (cys). The apparent conservation of these properties during evolution suggests that they play an important role in protein structure.

Many other properties of the amino acids have been proposed and studied (see the extensive listing in Nakai, Kidera, & Kanehisa 1988). Reasonable groupings of amino acids that are slight variants of Dayhoff's exchange groups have been investigated (Sander & Schulz 1979; Taylor 1986). Many alternative scales have been explored for properties such as volume (Zimmerman, Eliezer, & Simha 1968; Chothia 1975; Levitt 1976) and hydrophobicity (Nozaki & Tanford 1971; Charton & Charton 1982; Rose *et al.* 1985; Cornette *et al.* 1987). Empirical measures of properties such as refractivity, isoelectric point, heat capacity, and partial specific volume have been considered. And statistical properties have been derived from protein-structure data, such as solvent accessibility (Chothia 1976), secondary-structure preference (Chou & Fasman 1978), and tertiary contact patterns (Ponnuswamy, Prabhakaran, & Manavalan 1980). Richardson and Richardson (1989) provide a good survey of the major properties of each amino acid and the roles they have been observed to play in protein structures.

Knowledge about the chemical and physical properties of amino acids has become crucial to many tasks in biology. Such knowledge is needed to facilitate the identification and modeling of new proteins by enhancing alignments to known proteins. Typically, related proteins are identified by aligning a new protein with the amino acid sequences of proteins in a protein database. A match is indicated when the alignment produces a score significantly higher than for random alignments (Doolittle 1981), and the alignment can then be used as a basis for constructing a model of the new protein (Blundell *et al.* 1987). However, sometimes proteins with similar structures do not have significantly high sequence identity (e.g. the wide variety of distinct beta-barrels; Chothia 1988). Still, because the structures are similar, we expect that there is some pattern to the amino acid sequences, at least reflecting structural constraints, independent of evolutionary relationships. Amino acid properties can be used to help alleviate this "hidden-homology problem" by enabling a more intelligent local matching function in alignments. The decision about whether to match two amino acids should be made to depend on their chemical and physical similarities, which determine whether they can play the same role in the protein structure (see the PIMA algorithm; Smith & Smith 1990).

Despite the plethora of information about the properties of the amino acids, our knowledge about the roles amino acids play in proteins remains fundamentally weak. Specifically, there is a great deal of uncertainty about which properties are most important for determining protein structure. Each of the amino acid side-chains consists of multiple properties, producing orthogonal dimensions of similarity, and there are many cases where different properties disagree on the similarity between two amino acids. For example, threonine is like valine in shape, and it is like serine in that it contains a hydroxyl; however, valine and serine are not similar to each other in either of the respects. Other questionable relationships include the potential hydrophobicity of phenylalanine (overshadowed by its aromaticity), the aromaticity of histidine (overshadowed by its positive charge), and the occasional circumstances in which cysteine plays a generic role as a small residue, rather than a specific role in an active site or disulfide-bridge. We cannot easily resolve these conflicts because we do not know how to weight the relevance of these overlapping properties (Sneath 1966). Studies have used correlational techniques (Grantham 1974) and factor analyses (Kidera *et al.* 1985) to try to extract an underlying basis set of similarities, typically concluding that some variations of bulk and hydrophobicity are the predominant properties in general. However, these generalizations do not adequately explain the wide diversity of amino acid substitution patterns observed in multiple alignments of proteins, and do not provide sufficient guidance in understanding what substitutions are acceptable at individual sites.

Context-Dependence

Our approach to dealing with the hidden-homology problem is based on the observation that the relevance of the amino acid properties depends on their context. Each amino acid side-chain occurs in some environment consisting of the local secondary and tertiary structure. These contextual factors have a significant influence on the rates of substitution between the amino acids. The acceptability of various amino acids at a site has been observed to correlate with the polarity of contacting chemical groups, as well as conserving hydrogen-bonding patterns (Warme & Morgan 1978). Ponnuswamy et al. (1980) found that the average level of hydrophobicity among surrounding residues (in 3D space) has a strong effect on the distribution of amino acids at a given site. Overington et al. (1992) collected independent substitution tables for observed amino acid replacements in α-helices versus β-sheets, and also for exposed versus buried sites, finding significant deviations from the generalized patterns of similarity. And Ouzounis et al. (1993) found specialized distributions of amino acid preferences based on the types of secondary structures against which a side-chain packs.

Of course, this secondary and tertiary context information may not be available if the structures of neither protein in the alignment are known. Although approaches such as 3D-1D profiling (Bowie, Luthy, & Eisenberg 1991) can use known protein structures to greatly enhance alignments, there are still many cases in which alignments must be done without the aid of structure information. Thus we need to refine our understanding of the amino acid properties to improve the quality and sensitivity of alignments based on sequence information alone.

Sequential Context

There are several ways in which the local sequence around each amino acid can be used to approximate the context of the site in the environment of the protein. Goldman et al. (1996) use the internal states of a hidden Markov model for recognizing secondary structure classes as conditions for estimating independent substitution rates. Our proposal is to use the *pattern of hydropathy among neighboring amino acids* in a contiguous window surrounding a site as a surrogate for context. Typically, some of these residues will make contact with the side-chain of the residue at the center of the site. For example, in β-sheets, neighbors $i \pm 2$ contact the central residue, and in α-helices, neighbors $i-3$ and $i-4$ in the previous loop and $i+3$, and $i+4$ in the subsequent loop often make contact with the central amino acid at position i (Schulz & Schirmer 1979). Residues that contact the side-chain of the central residue at a site participate in its tertiary environment, and because of their sequential locality, they usually participate in the same secondary structure as well. Therefore, on average, these neighboring residues should reflect the nature of the surrounding environment, and we should look for patterns within such a window to estimate the context for the central amino acid. Ultimately, we want to use this sequential definition of context to distinguish the relevance of amino acid properties in different situations.

Method

Overview of Experiment In this section, we describe an experiment that explores the utility of sequential context in restricting the use of amino acid properties in alignments. We will present a specific definition of context, and we will allow amino acids to be matched together independently in each context according to one or a combination of a pre-defined (manually chosen) set of properties. The selection of different properties in each context can be treated as hypotheses. To evaluate these hypotheses we will apply them to re-represent example amino acid sequences with symbols based on their hypothesized properties, which will affect their alignments, and then measure the increase or decrease in the sensitivity of recognizing when two proteins are in the same fold-class based on significance of alignment score. We will use a stochastic hill-climbing algorithm to search for the best combination of properties in each context, as evaluated over a training set of proteins, and we will show that after 1000 iterations, the sensitivity of alignments increases on a separate testing set. For comparison, we will also show that this approach even does better than using Dayhoff's (1978) MDM78 substitution table. Our goal is to show that sequential context can allow the expression of finer relationships among the amino acids by capturing independent dimensions of similarity in different situations.

Data In these experiments, we used a set of 199 proteins, representing a range of α, β, $\alpha+\beta$, and α/β structures (see (Ioerger 1996) for a list of the proteins, including their chain identifiers and sequence limits). The amino acid sequences were obtained from the Brookhaven Protein Databank (PDB; Bernstein *et al.* 1977). They ranged in length from 39 residues (4gcr) to 478 residues (2taa). These proteins have been assigned to 37 distinct fold classes by visual inspection (Pascarella & Argos 1992), with at least two proteins per fold. Many proteins in the same fold class had low homology: over 50% of such pairs had less than 25% sequence identity.

Alignment Algorithm The alignment algorithm we used is a variant of the linear-space global-alignment algorithm by Myers and Miller (1988). Gap weights were chosen by a one-time optimization in which a grid-based search was performed to identify parameters that produced the maximal average Z-score (see below) on our dataset. The selected weights were:

```
- Ala - Gly - Asp - LEU - Arg - Ile - Phe -

  phob  phil  phil        phil  phob  phob

   0     1     1           1     0     0
```

Figure 1: Example computation of sequential context. In this example, we will compute the context for the central residue (leucine) in the peptide fragment above (line 1). First, all six neighbors (three on either side) are mapped to their hydrophobicity values ('phil' for hydrophilic, 'phob' for hydrophobic; line 2). Then the pattern is written as a bit-vector (1 for hydrophilic, 0 for hydrophobic; line 3). The context number is generated by interpreting the bit-vector as a binary number (LSB-first): $011100_2 = 14$.

gap-open-penalty$= -1.5$ and gap-extension-penalty$= -0.2$, relative to a match value of 1.0 for identical residues. The final alignment scores (sum of matches minus gap penalties) were normalized by dividing by the length of the shorter sequence to make the degree of matching independent of length (like a percentage). The algorithm was implemented in C using threads on a shared-memory multi-processor SGI Power Challenge. The parallelism was at the level of individual alignments, distributing each pair of proteins to be aligned to an independent processor.

Definition of Context As discussed above, our goal is to explore the utility of context defined in terms of sequence information alone. In this experiment, we chose a window size of seven to capture potential side-chain contacts in α-helices and β-sheets. Such a window contains six neighbors for each site: $i-3$, $i-2$, $i-1$, $i+1$, $i+2$, and $i+3$. Within this window, we looked at the pattern of hydrophobicity, since the local hydrophobicity of an environment is known to have a significant effect on substitution patterns (Ponnuswamy, Prabhakaran, & Manavalan 1980). We divided the 20 amino acids into two subsets:

$$\text{hydrophobic}=\{A,F,I,L,M,P,V,W\} \text{ and}$$
$$\text{hydrophilic}=\{C,D,E,G,H,K,N,Q,R,S,T,Y\}.$$

Each of the six neighbors in the window was assigned a 0 if it was hydrophobic or a 1 if it was hydrophilic.[1] This produced a six-bit pattern which, when interpreted as an integer, resulted in a number between 0 and 63. So this definition distinguishes 64 unique contexts (two states for each of six neighbors: $2^6 = 64$). An illustration of computing the context at a hypothetical site is shown in Figure 1.

[1]The few amino acids with undefined neighbors near the termini of sequences can be assigned special context identifiers to force them to match strictly by identity during sequence alignments.

Table 1: The 18 properties used in these experiments, expressed as subsets of amino acids (in 3-letter codes) that have each property.

property	*amino acids*
core-hydrophobic	met ile leu val
small-hydrophobic	met ile leu val ala pro
large-hydrophobic	met ile leu val phy trp
all-hydrophobic	met ile leu val ala pro phe trp
charged	asp glu his arg lys
small-polar	ser thr gln asn
small-strong-polar	ser thr gln asn asp glu
strongly-polar	gln asn asp glu his arg lys
all-polar	ser thr gln asn asp glu his arg lys
amide	gln asn
carbonyl	asp glu gln asn
positive	his arg lys
positive-no-his	arg lys
aromatic	phe trp tyr
left-handed	gly asn
tiny	ala cys gly ser
hydroxyl	ser thr
carboxyl	asp glu

Amino Acid Properties In each context, we want to allow amino acids to match according to a unique set of properties. We express properties in terms of discrete subsets of amino acids that contain each property (Taylor 1986), as opposed to using continuous scales. We have selected a set of 18 properties that are expected to be relevant to protein structure (Table 1). Many of these properties are variants of Dayhoff's exchange groups, typically corresponding to greater or lesser degrees of some core property (e.g. the series of groups based on polarity: positive $\subset$ charged $\subset$ strongly-polar $\subset$ all-polar). Other properties are based on constituent groups (e.g. hydroxyl, carboxyl), and 'left-handed' refers to flexibility in backbone torsion angles.

Hypothesis Representation A hypothesis consists of a specification of which properties are relevant in each context, which is effectively a theory about the context-dependence of the amino acid properties. A hypothesis is represented as a *list of 64 partitions* of the 20 amino acids, one partition for each context. Initially, all amino acids (in all contexts) are assigned to singleton classes, treating them all as unique. But groups of amino acids can be merged into classes in the partition for any context, based on some property that they have in common. A wide variety of partitions of amino acids can be generated by a performing a sequence of merging operations according to multiple properties in various orders. Figure 2 illustrates this process of merging amino acids in partitions.

A hypothesis about amino acid properties in this form may be applied in the following way to influence local matching during sequence alignment. When considering whether to match the amino acids at position i in sequence 1 and position j in sequence 2, first the context in the window surrounding each amino acid computed (as in Figure 1). If the contexts are not equal, then the amino acids are matched solely on the basis of identity. However, if the amino acids occur in the same context, then the corresponding partition for that context is looked up in the hypothesis table. If the two amino acids both belong to the same class in the partition, then they are counted as a match (scoring 1). Otherwise, they are counted as a mis-match (scoring 0). The initial hypothesis, with 64 singleton partitions, results in generating identity alignments, since all amino acids are in distinct classes and will only be matched when they are identical, regardless of context.

```
a. ((A)(C)(D)(E)(F)(G)(H)(I)(K)(L)
   (M)(N)(P)(Q)(R)(S)(T)(V)(W)(Y))

b. ((ACGS)(D)(E)(F)(H)(I)(K)(L)(M)
   (N)(P)(Q)(R)(T)(V)(W)(Y))

c. ((APMILV)(CGS)(D)(E)(F)(H)(K)
   (N)(Q)(R)(T)(W)(Y))
```

Figure 2: Example of merging amino acids by properties in partitions. a) The initial partition, in which each amino acid is in a unique class. b) The same partition as (a) after merging amino acids according to the property 'tiny'. c) The same partition as (b) after merging amino acids according to the property 'small-hydrophobic'. Note that alanine (A) has both properties, but is extracted from the the first group and placed in the second, due to the order in which these merges were performed.

Evaluation: Average Z-scores We can evaluate and compare various hypotheses by applying them to alignments within a set of proteins and measuring the effect on the *sensitivity* of detecting structurally-related proteins. Proteins may be recognized as belonging to the same fold class when they have a significantly high alignment score. Molecular biologists often measure the significance of an alignment score by comparing it to a background distribution of alignment scores between unrelated proteins (with different folds) and computing the statistical Z-score (Doolittle 1981). The mean μ and the standard deviation σ of scores in this distribution are determined, and the Z-score for an alignment score sc is then computed by: $Z = (sc - \mu)/\sigma$. Z-scores above a cutoff in the range of 3.0 to 6.0 are generally assumed to indicate a structurally-related pair of proteins, since the probability that they came from the background distribution of unrelated proteins is exceedingly small. The sensitivity of an alignment procedure (using a particular hypothesis about amino acid properties) may thus be quantified over a particular dataset via the *average Z-score* for all pairs of proteins in the same fold-class. This computation requires a complete table of all pairwise alignment scores within the set of proteins to be computed, including non-same-fold pairs for background distributions, which consists of nearly 20,000 comparisons for our set of 199 proteins. Computing this alignment-score table, which must be re-done for every hypothesis that is tested, takes about 30 seconds on an SGI Power Challenge using three processors.[2]

Search Procedure This experimental setup allows us to explore how the relevance of amino acid properties depends on their context. While we have an idea *a priori* about what properties might be useful, we do not know the contexts in which they will be most appropriate. Thus we will use a search procedure to incrementally generate and test new hypotheses. We will start with the initial hypothesis that all amino acids are unique, which is equivalent to constructing alignments by identity. Then a random property among the 18 listed in Table 1 and a random context among the 64 possibilities will be chosen. The current hypothesis (list of 64 amino-acid partitions, one for each context) will be updated by merging amino acids (as in Figure 2) in the partition for the selected context according to the selected property. The pairwise table of alignment scores will be re-computed, and the change in Z-score will be calculated. If the Z-score increases, then the updated hypothesis will be kept for the next iteration; otherwise the previous hypothesis will be restored. Thus this search procedure is essentially performing a stochastic form of hill-climbing.

Cross-Validation In our dataset of 199 proteins, there are 950 *pairs* of proteins that belong to the same fold class. These were divided into 10 balanced, disjoint subsets. For each division, 90% of the pairs were used for training (to guide the selection properties in contexts), and the remaining 10% of pairs were used to evaluate the change in average Z-score. This cross-validation approach was taken to avoid over-fitting by ensuring that the pairs of proteins used to monitor the improvement were different from the pairs of proteins used to guide the search.

Results

Improvement by Search We ran the search process described above for 1000 iterations. The resulting average Z-scores for various methods, computed on independent test sets for each run of cross-validation, are shown in Table 2. First, notice that the search was consistently able to improve the sensitivity of sequence alignments by tuning partitions of amino acids to each context. For example, in the first run, the average Z-score for identity alignments on the test set (10% of the same-fold pairs of sequences) was 5.41. After 1000 iterations of search, which were guided by evaluating average Z-scores on the remaining 90% of the data, the average Z-score was again evaluated on the test set, and was found to have increased to 5.95. This means that, on average, alignments based on context-dependent properties had higher Z-scores, and thus became easier to distinguish from the background distribution of scores from alignments of unrelated proteins (i.e. lowered the probability of making the mistake of predicting that some other protein is more related).

Table 2: Average Z-scores on independent test sets for each run of cross-validation. 'identity' means alignments were done by identity. 'search' means alignments were doing using the best re-representation found (based on the training data) during 1000 iterations of search. 'mdm78' means alignments were done using Dayhoff's substitution table. 'search–mdm78' gives the relative improvement of the best representation found during search, compared to using the substitution table method.

run	identity	search	mdm78	search–mdm78
1	5.40963	5.94772	6.16741	–0.21969
2	6.59134	7.10707	6.7676	+0.33947
3	5.24691	5.74617	5.69068	+0.05549
4	4.19741	4.70337	4.50076	+0.20261
5	5.2671	5.77199	5.91217	–0.14018
6	4.87253	5.34408	5.24804	+0.09604
7	4.41745	4.87344	4.59071	+0.28273
8	5.105	5.68543	5.56681	+0.11862
9	5.08586	5.53036	5.25719	+0.27317
10	5.96101	6.62923	6.27041	+0.35882
avg	5.21542	5.73389	5.59718	+0.13671

Figure 3 shows the increase in average Z-score, averaged over all 10 runs, as the search progressed. The data points in the graph are the average Z-scores, averaged over all 10 runs, for the best hypothesis discovered up to that point in the search, measured every 100 iterations on the independent test sets. Since the starting hypothesis (at iteration 0) was equivalent to the 'null' representation that treats amino acids in all contexts as unique, the baseline for the curve (lower horizontal line on the graph) corresponds to the average Z-score for alignments using amino acid identities. This score was computed to be 5.215, averaged over all 10 runs. After 1000 iterations of search, the average Z-score, averaged over all 10 runs of cross-validation, had risen by +0.519 to 5.734. Again, this increase in average Z-scores means that proteins with similar structures are easier to recognize on the basis of sequence alignments.

[2]Computational resources were provided by the National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign.

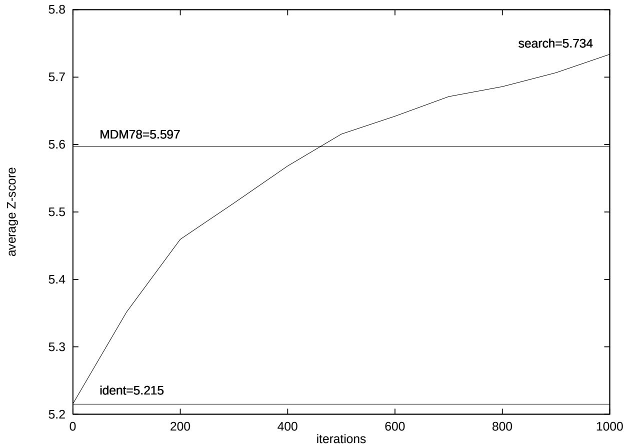

Figure 3: Improvement in average Z-score during search. This curve represents the average Z-score, evaluated every 100 iterations on the independent test sets, averaged over all 10 runs of cross-validation.

Comparison to MDM78 To assess the *importance* of this improvement in sensitivity, we also computed the average Z-score for alignments constructed using the MDM78 substitution table, which is one of the most widely used methods for improving sequence alignment algorithms (Dayhoff, Schwartz, & Orcutt 1978). The average Z-scores on the test sets for each run of cross-validation are also shown in Table 2.[3] The MDM78 method produced an average Z-score of 5.597 over all 10 runs, an increase of +0.382 (shown by the upper horizontal line in Figure 3). Our search procedure produced a higher average Z-score than MDM78 on most, but not all, runs of cross-validation. Based on a paired t-test, we can state that the improvement by our search-based method was generally greater than the improvement by Dayhoff's MDM78 substitution table method, at a confidence level of $p < 0.06$.

Resulting Hypotheses Table 3 presents the resulting partitions for eight out of the 64 contexts after 1000 iterations of search (for the tenth run of cross-validation, which demonstrated the maximum improvement over MDM78: +0.359). In particular, the table lists partitions for all eight contexts in which the three C-terminal neighbors of a site are all hydrophobic, and the three N-terminal neighbors have arbitrary combinations of hydropathy states. These partitions illustrate that a great deal diversity exists in the substitution constraints in different contexts. By running the search for many more iterations, and perhaps introducing new operators for swapping amino acids among classes, the partitions might converge upon unique groupings of amino acids that are most useful for aligning distantly-related sequences. It is also interesting to note some potential correlations among the groupings of amino acids in different contexts. For example, the first two partitions are fairly similar, and they belong to contexts in which only the hydropathy of neighbor $i-3$ is different. This suggests that higher level patterns may exist that could be explored by representing contexts in a more flexible description language that can be used to group similar contexts together so that they can share the same partition.

[3]To allow a fair comparison, we re-optimized the gap-weights for the substitution-table method: gap-open-penalty=–120, gap-extension-penalty=–8.

Table 3: Eight of the 64 partitions after 1000 iterations of search in the tenth run of cross-validation. The contexts are listed as bit-patterns, which indicate the hydropathy (0=hydrophobic, 1=hydrophilic) of neighbors, reading left to right as $i-3$ to $i+3$. These partitions illustrate the variety of relevant properties selected in different contexts. The full hypothesis discovered by search had partitions for all 64 contexts.

context: *partition*
000000: (AILMPV)(CS)(DEHKQR)(FW)(GN)(T)(Y)
100000: (AILMPV)(C)(D)(E)(FW)(GN)(H)(K)(Q)(R)(S)(T)(Y)
010000: (A)(C)(DE)(FWY)(G)(HKR)(I)(L)(M)(NQ)(P)(ST)(V)
110000: (ACG)(DEHKNQRST)(FILMVW)(P)(Y)
001000: (ACGS)(DENQ)(F)(HKR)(I)(L)(M)(P)(T)(V)(W)(Y)
101000: (AILMPV)(C)(DE)(FW)(G)(HKR)(N)(Q)(S)(T)(Y)
011000: (AILMPV)(C)(DENQ)(F)(G)(HKR)(S)(T)(W)(Y)
111000: (AP)(C)(DE)(FILMVW)(G)(H)(K)(N)(Q)(R)(S)(T)(Y)

Dependence on Homology An important question is whether the new context-based representation discovered by the search gives more improvement to alignments that have high homology or low homology. Specifically, it is important to improve the significance of low-homology alignments ($Z < 3.0$), since these are the ones that cause the most difficulty in homology modeling. Figure 4a shows the dependence of the improvement in Z-score on the original Z-score based on alignments by identity. While not all sequences receive increased Z-scores, it appears that many alignments improve in significance by a few standard deviations. This trend appears fairly uniform with respect to the original Z-score, so the method is able to improve low-homology alignments as well as high-homology alignments. When the Z-score happens to be decreased by this new representation, the loss of significance is not very great. For comparison, Figure 4b shows the de-

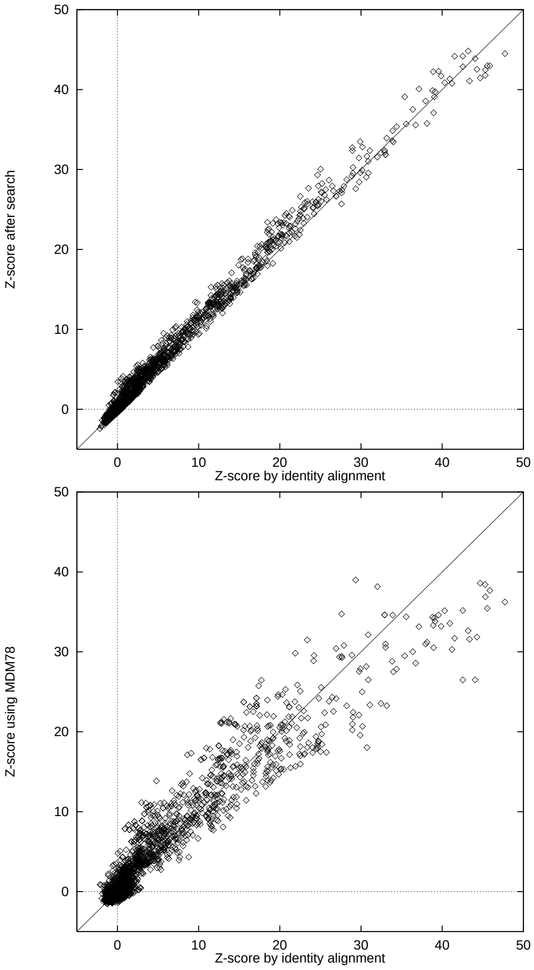

Figure 4: Dependence of improved Z-scores on original Z-scores. a) Context-dependent representation discovered after 1000 iterations of search. b) MDM78 substitution table.

pendence of the improvement in Z-score when using the MDM78 substitution table, as a function of the Z-score when doing alignments by amino acid identities. The variance is considerably larger; the Z-score for some alignments is greatly increased, and for other alignments it is greatly decreased.

Effect on Alignments To get a feel for how this new context-dependent representation actually affects sequence alignments, we can look at both the new matches introduced into particular alignments, as well as any shifts in the location of gaps. To illustrate, we selected an alignment with one of the greatest increases in Z-score: between the N-terminal half of the heavy chain of the Fab fragment of human IgG (2fb4) and the N-terminal half of the light chain of the Fab fragment of mouse IgA (2fbj). These are immunoglobulins with a beta-sandwich structure (two beta-sheets packed face-to-face). Figure 5 presents the alignment based on identities, the alignment based on our new representation, and an alignment constructed manually based on structural correspondences.

In this case, the new representation adds only three new matches. However, these new matches are enough to cause shifts in the location of gaps, resulting in a different alignment. The alignment score for identities is 0.153, which has a Z-score of 0.534. Using our context-dependent representation, the alignment score increases to 0.225, and the Z-score increases by over 3 standard deviations to 3.84. The alignment using our context-dependent representation seems slightly better than the identity alignment (closer to the structural alignment) in the first third of the protein, slightly worse in the middle, and just as bad in the last third. Formal measurement of the effect of such re-representations on sequence alignments remains for future work. However, it is clear that alternative representations will have to introduce many more new matches to improve the quality of alignments in addition to their significance.

Discussion

The hypotheses generated by each of the searches in this experiment consist of a listing of what groupings of amino acids are most relevant for each of the 64 contexts we defined, which effectively represents a context-dependent theory about the relevance of amino acid properties. In the evaluation of these searches, we showed how a representation such as this could be used to improve a sequence alignment algorithm by allowing new matches between similar but non-identical amino acids, conditioned on their sequential neighbors.

Our syntactic approach to defining context purely in terms of sequential neighbors of each site allows us to refine our knowledge of amino acid substitution patterns without having to know the structures of the proteins involved. This is important for a variety of applications, such as increasing the sensitivity of sequence database searches (usually accomplished with substitution tables), constructing multiple alignments of families of proteins whose structures are currently unknown (the vast majority of cases), or analyzing the phylogenetic relationships among such a family of proteins, which relies heavily on an accurate assessment of the similarities between molecular sequences. Clearly, knowledge of at least one of the protein structures involved would greatly facilitate making decisions about the appropriateness of amino acid matches, as in methods like 3D-1D profiling (Bowie, Luthy, & Eisenberg 1991).

There are many ways to extend the experiment described in this paper. First, our definition of context was both limited and static. The size of the window of neighbors could be extended, and other properties besides the hydophobicity of these residues could be considered. Furthermore, it might be possible to group

```
a. identity alignment

  EVQLVQSGGGVVQPGRSLRLSCSSSGFIFSSYAMYWVRQAPGKGL
  |  | ||                |     ||   |
  EIVLTQSPAITAASLGQKVTITCSASSSVSSLHWYQQKSGTSPKP

  EWVAIIWDDGSDQHYADSVKGRFTISRNDSKNTLFLQMDSLRPED
      |          |  |  ||  |   |     |       ||
  WIYEISKL-------ASGVPARFSGSG--SGTSYSLTINTMEAED

  TGVYFCARDGGHGFCSSASCFGPDYWGQ
     | |           |
  AAIYYCQQWTYPLITFGAGTKLELK---


b. alignment with context-dependent partitions

  EVQLVQSGGGVVQP-GRSLRLSCSSSGFIFSSYAMYWVRQAPGKG
  |  | ||        |  *   || |    |     |  |  |
  EIVLTQSPAITAASLGQKVTITCSASSSVSSLH---WYQQKSGTS

  LE-WVAIIWDDGSDQHYADSVKGRFTISRNDSKNTLFLQMDSLRP
   * |   |    |       |  ||  |   |     |  *
  PKPWIYEISKLASG------VPARF--SGSGSGTSYSLTINTMEA

  EDTGVYFCARDGGHGFCSSASCFGPDYWGQ
  ||   | |           |
  EDAAIYYCQQWTYPLITFGAGTKLELK---


c. structural alignment

  EVQLVQSGG-GVVQPGRS-LRLSCSSSGFIFSSYAMYWVRQAP--
  |  | ||        |         ||             |
  EIVLTQSPAITAASLGQKVTITCSASSSV-----SSLHWYQQKSG

  GKGLEWVAIIWDDGSDQHYADSVKGRFTISRNDSKNTLFLQMDSL
       |   |         |  || ||  |         |
  TSPKPWIYEI------SKLASGVPARFSGSGSG--TSYSLTINTM

  RPEDT-GVYFCARDGGHGFCSSASCFGPDYWGQ
    ||    |                 |
  EAEDAAIYYCQQWT----YPLITFGAGTKLELK
```

Figure 5: Effect of the context-dependent representation on the alignment of two immunoglobulins: 2fb4-H (top sequence) and 2fbj-L (bottom sequence). The top alignment was done with identities. Vertical bars indicate identity matches. The second alignment was done with our improved representation. The asterisks indicate matches based the context-dependent partitions, rather than identity. The third alignment was constructed manually from comparison of the structures.

some of these contexts together, for example by clustering, to merge rules for sites with similar substitution patterns. In the most general case, context could be described as arbitrary patterns of properties among an arbitrary set of neighbors using expressions in a general concept-description language.

In addition to the definition of context, we could have manipulated the set of properties used in the search. New groups of amino acids could be formed by swapping amino acids from the group in which they are most likely to reside to other groups to which they could plausibly belong. This suggests a hypothesis-testing approach in which secondary properties, such as the hydrophobicity of tryptophan and phenylalanine, are used to intelligently guide the generation and evaluation of alternative partitions in random contexts. This approach is not only applicable to qualitative descriptions of amino acid properties in terms of subsets of amino acids, but could also be used with numeric scales, such as by identifying the optimal contexts for the various measures of hydrophobicity and bulk.

The power of our search-based approach comes from an effective combination of both domain expertise and computational resources. Understanding the forces involved in determining protein structure is an extremely difficult task. The growing databases of protein sequence and structure provide an excellent source of examples of this relationship, and we want to use computational techniques to extract the patterns implicit in this data. The search method we have presented in this paper incrementally explores a space of these patterns looking for those that fit the data well. But the search must be adequately controlled to be both effective and efficient. This control is exercised through the use of background knowledge about the kinds of patterns that are expected (i.e. definitions of contexts and properties). This knowledge-based approach contrasts with other methods, such as using Bayesian techniques to calculate prior weights for mixtures of similarity scales that best explain a set of amino acids observed at a site (Brown *et al.* 1993).

In our experiments, we used biochemical knowledge to choose a particular definition of context that we believed would be useful in separating sites with distinct substitution patterns, and to choose a particular set of properties that we expected might be relevant. These choices resulted in the discovery of representations of amino acid sequences that make significant improvements in alignment sensitivity within 1000 iterations, reflecting sufficient constraint of the search. However, each iteration is very computationally intensive, requiring tens of thousands of sequence alignments, and the entire experiment accumulated approximately one day of CPU time on a supercomputer. Therefore, we are near pragmatic limits on the computational complexity of the search, which emphasizes the importance of the background knowledge we used. Without these initial guesses, the search for relevant groups of amino acids in arbitrary contexts would be too unconstrained. To make additional gains in sequence alignment sensitivity by this approach will require using more background knowledge to help make intelligent decisions about how to explore the space of hypotheses about the context-dependence of the relevance of amino acid properties.

References

Bernstein, F.; Koetzle, T.; Williams, G.; Meyer, E.; Brice, M.; Rodgers, J.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112:535–542.

Blundell, T.; Sibanda, B.; Sternberg, M.; and Thornton, J. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326:347–352.

Bowie, J.; Luthy, R.; and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170.

Brown, M.; Hughey, R.; Krogh, A.; Mian, I.; Sjolander, K.; and Haussler, D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 47–55.

Charton, M., and Charton, B. 1982. Structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology* 99:629–664.

Chothia, C. 1975. Structural invariants in protein folding. *Nature* 254:304–308.

Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology* 105:1–14.

Chothia, C. 1988. The fourteenth barrel rolls out. *Nature* 333:598–599.

Chou, P., and Fasman, G. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology* 47:45–148.

Cornette, J.; Cease, K.; Margalit, H.; Spouge, J.; Berzofsky, J.; and DeLisi, C. 1987. Hydrophobic scales and computational techniques for detecting amphipathic structures in proteins. *Journal of Molecular Biology* 195:659–685.

Dayhoff, M.; Eck, R.; and Park, C. 1972. A model of evolutionary change in proteins. In Dayhoff, M., ed., *Atlas of Protein Sequence and Structure*, volume 5. National Biomedical Research Foundation: Silver Springs, MD.

Dayhoff, M.; Schwartz, R.; and Orcutt, B. 1978. A model of evolutionary change in proteins. In Dayhoff, M., ed., *Atlas of Protein Sequence and Structure*, volume 5 (supplement 3). National Biomedical Research Foundation: Silver Springs, MD. 345–358.

Doolittle, R. 1981. Similar amino acid sequences: Chance or common ancestry? *Science* 214:149–159.

Goldman, N.; Thorne, J.; and Jones, D. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Biology* 263:196–208.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.

Ioerger, T. 1996. *Change-of-Representation in Machine Learning, and an Application to Protein Structure Prediction.* Ph.D. Dissertation, University of Illinois, Department of Computer Science.

Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; and Scheraga, H. 1985. Statisitical analysis of the physical properties of the 20 naturally occuring amino acids. *Journal of Protein Chemistry* 4:23–54.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology* 104:59–116.

Myers, E., and Miller, W. 1988. Optimal alignments in linear space. *CABIOS* 4:11–17.

Nakai, K.; Kidera, A.; and Kanehisa, M. 1988. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering* 2:93–100.

Nozaki, Y., and Tanford, C. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. *Journal of Biological Chemistry* 246:2211–2217.

Ouzounis, C.; Sander, C.; Scharf, M.; and Schneider, R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness. *Journal of Molecular Biology* 232:805–825.

Overington, J.; Donnelly, D.; Johnson, J.; Sali, A.; and Blundell, T. 1992. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Science* 1:216–226.

Pascarella, S., and Argos, P. 1992. A data bank merging related protein structures and sequences. *Protein Engineering* 5:121–137.

Ponnuswamy, P.; Prabhakaran, M.; and Manavalan, P. 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochimica et Biophysica Acta* 623:301–316.

Richardson, J., and Richardson, D. 1989. Principles and patterns of protein conformation. In Fasman, G., ed., *Prediction of Protein Structure and the Principles of Protein Conformation.* Plenum Press: New York. 1–98.

Rose, G.; Geselowitz, A.; Glenn, J.; Lee, R.; and Zehfus, M. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* 229:834–838.

Sander, C., and Schulz, G. 1979. Degeneracy of the information contained in amino acid sequences: Evidence from overlaid genes. *Journal of Molecular Evolution* 13:245–252.

Schulz, G., and Schirmer, R. 1979. *Principles of Protein Structure.* Springer-Verlag: New York.

Smith, R., and Smith, T. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Biochemistry* 87:118–122.

Sneath, P. 1966. Relations between chemical structure and biological activity in peptides. *Journal of Theoretical Biology* 12:157–195.

Taylor, W. 1986. The classification of amino acid conservation. *Journal of Theoretical Biology* 119:205–218.

Warme, P., and Morgan, R. 1978. A survey of amino acid side-chain interactions in 21 proteins. *Journal of Molecular Biology* 118:289–304.

Zimmerman, J.; Eliezer, N.; and Simha, R. 1968. The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology* 21:170–201.