The Context-Dependence of Amino Acid Properties

Thomas R. Ioerger
Department of Computer Science
Texas A&M University

Homology Modeling

1. Identify a similar protein in a database.
   - string searching (e.g. BLAST)
   - local alignment (with gaps)
2. Assess statistical significance.
   - Compute Z-score.
   - If $>$ 30% identity, then similar structures (Sander & Schneider)
3. Global alignment/threading to match new amino acid segments to fragments of known structure.
4. Refine model using energy minimization...

Hidden-Homology Problem

- Proteins with similar structures do not always have similar sequences.

- E.g. at least 14 distinct beta-barrels, with various functions.

- Mandelate Racemase and Muconate Lactonizing Enzyme: 26% identity, 1.3$\AA$ RMS.

- Convergence, or just high divergence?

- "Twilight zone": $15 - 30\%$ identity, cannot prove *dis*-similarity.

Amino Acid Properties

Physical and chemical properties define *role* of each amino acid in a protein structure, and therefore what can *substitute* for it.

Want to consider these in alignments (PIMA).

Early evidence, Dayhoff et al. (1972): uneven substitution patterns, identified *exchange groups*.

Physical: volume, length, side-chain flexibility, backbone constraint, branching structure.

Chemical: hydrophobicity, polarity, charge, reactivity, aromaticity, hydrogen-bond donor/acc.

Statistical: solvent accessibility, secondary structure preference, tertiary contact patterns.

Overlapping Properties

Is Thr more like Val or Ser? Is His more like Phe or Arg? Role of Cys?

Orthogonal dimensions of similarity; Which are predominant?

Factor analyses: *bulk* and *hydrophobicity*.

Dependence on Context

1. Secondary structure and solvent access (Luthy, McLachlan, & Eisenberg, 1991)

2. Hydrophobicity of 3D environment (Ponnuswamy et al., 1980)

3. Chemical groups making tertiary contact (Warme and Morgan, 1978)

4. Secondary structure packed against (Ouzounis et al., 1993)

Our Approach

Goal: use context to determine which properties are relevant for assigning match scores between amino acids.

Assume no structure information is available.

Sequential Definition of Context:
pattern of hydrophobicity among $i-3 \ldots i+3$

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

2 states, 6 neighbors $\rightarrow 2^6 = 64$ contexts

Ala	Gly	Asp	LEU	Arg	Ile	Phe
0	1	1		1	0	0

$011100_2 = 14$

Hypotheses

represent amino acid similarities via *partitions*

{{MILV}{C}{D}{E}{G}{H}{K}{N}{P}{Q}{R}{S}{T}{FWY}}

allow a unique partition for each context

hypothesis ::= set of 64 partitions

initial hypothesis = singletons, identity matches

Merge Operator

for a given property, get amino acid subset

e.g. positive = {His,Arg,Lys}

form new class in partition

{{MILV}{C}{D}{E}{G}{N}{P}{Q}{S}{T}{FWY}{HRK}}

subsets defined for 18 typical properties

Experiment

Hill-climbing search:

```
choose random property (1 of 18)
   and random context (1 of 64)
merge amino acid subset in partition
evaluate; keep only if improved
iterate 1000 times
```

Evaluation: compute *average Z-scores*

- Align each pair of similar proteins; normalize by distribution of scores between unrelated proteins ($Z=(\text{score}-\mu)/\sigma$).

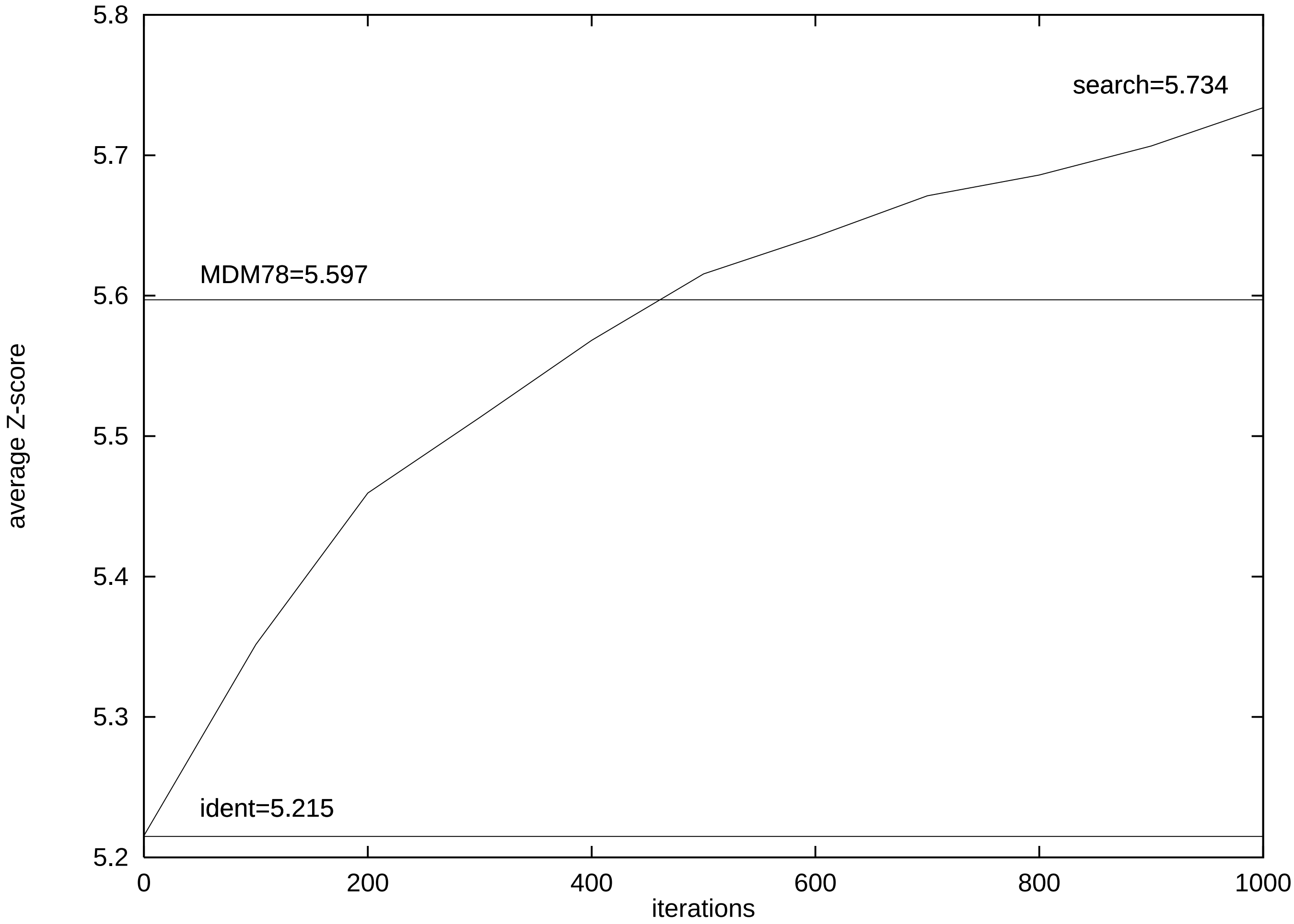
- measures *sensitivity* to detect relationships

Data: 950 pairs of proteins with similar structures (35 fold classes; Pascarella and Argos)

Results

identities	$Z = 5.22$
search	$Z = 5.73$*
MDM78	$Z = 5.60$

* on test sets, using 10-fold cross-val.



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)

Work in Progress

1. *Scales* for properties, instead of subsets.

2. Improving *quality* of alignments (RMS).

3. Analyzing structural alignments to bypass random search.

4. Tertiary context from known structure.

Conclusion

Property-based amino-acid matching should be conditioned on context.