A protocol for evaluating local structure alphabets

Rachel Karchin, Richard Hughey, Kevin Karplus

karplus@soe.ucsc.edu

Center for Biomolecular Science and Engineering

University of California, Santa Cruz

Outline of Talk

- What is a local structure alphabet?
- Example alphabets.
- What makes an alphabet good?
- Evaluation protocol.
- Results for several alphabets.

What is a local structure alphabet?

- Captures some aspect of the structure of a protein.
- Discrete classification for each residue of a protein.
- Easily computed, unambiguous assignment for known structure.
- Often based on backbone geometry or solvent accessiblity.

Backbone alphabets

Our first set of investigations was for a sampling of the many backbone-geometry alphabets:

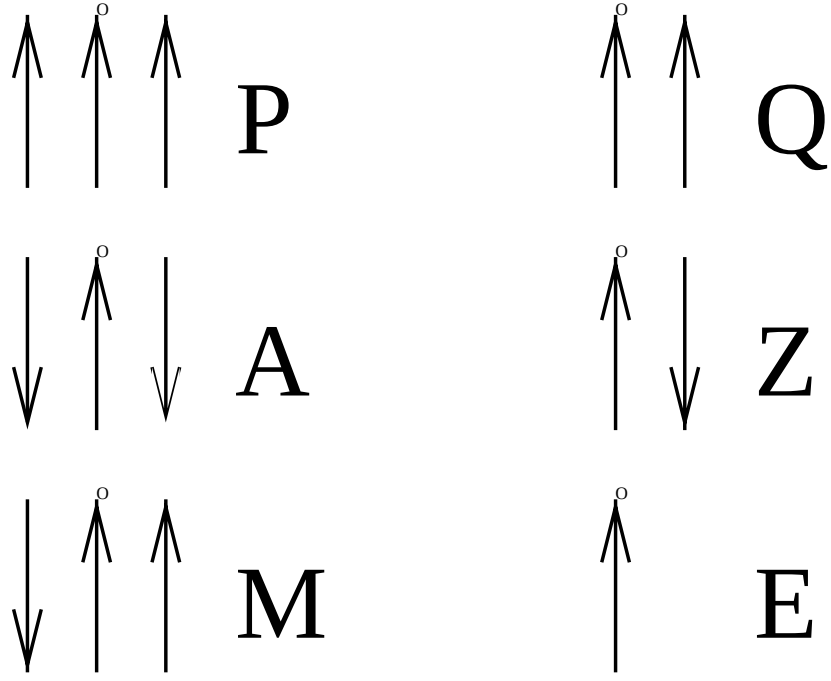
- DSSP
- our extensions to DSSP
- STRIDE
- DSSP-EHL and STRIDE-EHL
- HMMSTR ϕ-ψ alphabet
- α angle
- TCO
- de Brevern's protein blocks

DSSP

- DSSP is a popular program to define secondary structure.
- 7-letter alphabet: EBGHSTL
  - E = β strand
  - B = β bridge
  - G = 3_{10} helix
  - H = α helix
  - I = π helix (very rare, so we lump in with H)
  - S = bend
  - T = turn
  - L = everything else (DSSP uses space for L)

STR: Extension to DSSP

- Yael Mandel-Gutfreund noticed that parallel and anti-parallel strands had different hydrophobicity patterns, implying that parallel/antiparallel can be predicted from sequence.
- We created a new alphabet, splitting DSSP's E into 6 letters:



STRIDE

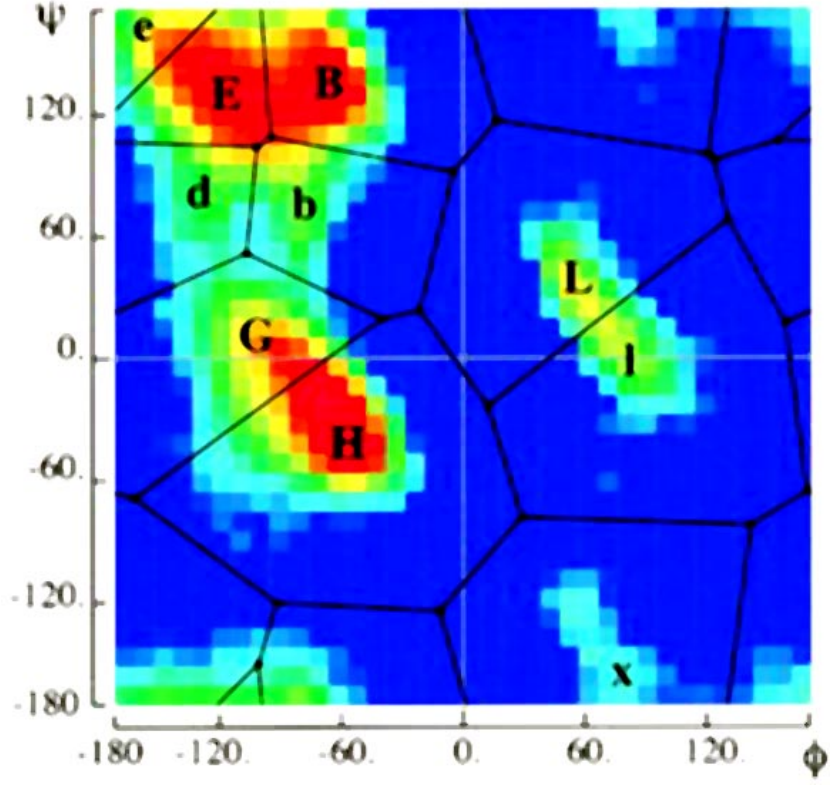
- A similar alphabet to DSSP, but uses more information in deciding classification for NMR and poor-resolution X-ray structures.
- 6-letter alphabet (eliminating DSSP's S=bend): EBGHTL
  - E = β strand
  - B = β bridge
  - G = 3_{10} helix
  - H = α helix
  - I = π helix (very rare, so we lump in with H)
  - T = turn
  - L = everything else

DSSP-EHL and STRIDE-EHL

- DSSP-EHL and STRIDE-EHL collapse the DSSP and STRIDE alphabets to 3 values
  - E = E, B
  - H = G, H, I
  - L = S, T, L
- The DSSP-EHL alphabet has been popular for evaluating secondary-structure predictors in the CASP and EVA experiments.

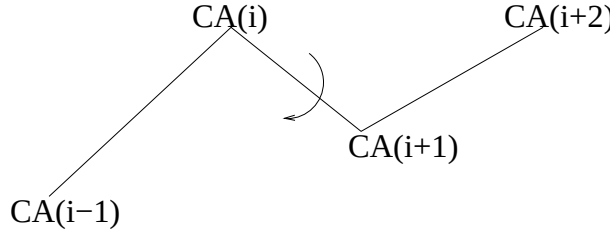
HMMSTR ϕ-ψ alphabet

- For HMMSTER, Bystroff did k-means classification of ϕ-ψ angle pairs into 10 classes (plus one class for cis peptides).
- We used just the 10 classes, ignoring the ω angle.

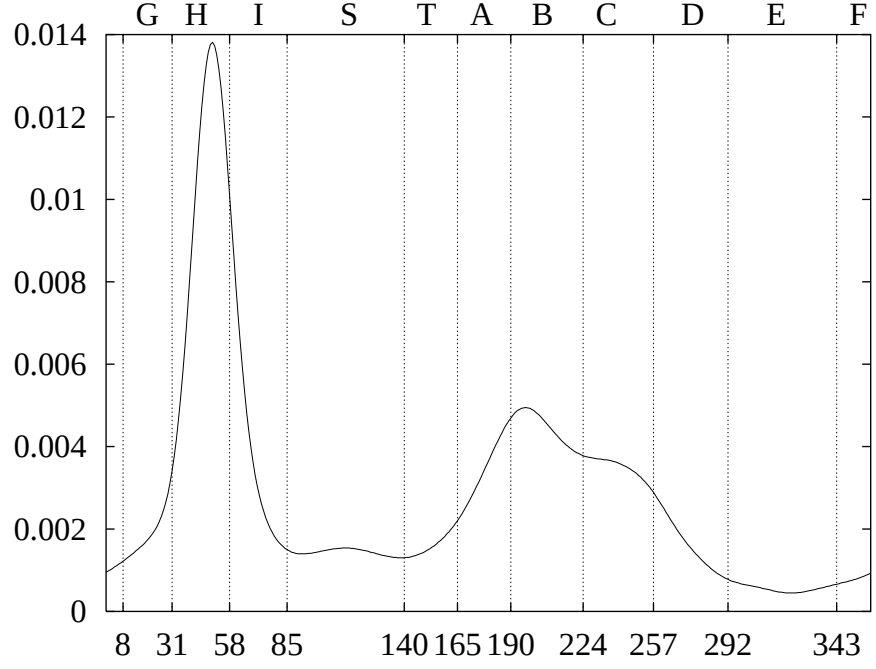


ALPHA11: α angle

- Backbone geometry can be mostly summarized with one angle per residue:

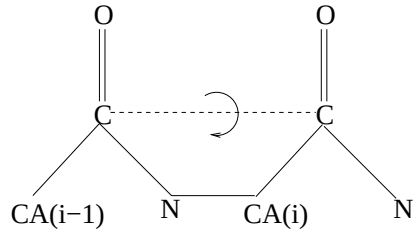


- We discretize into 11 classes:

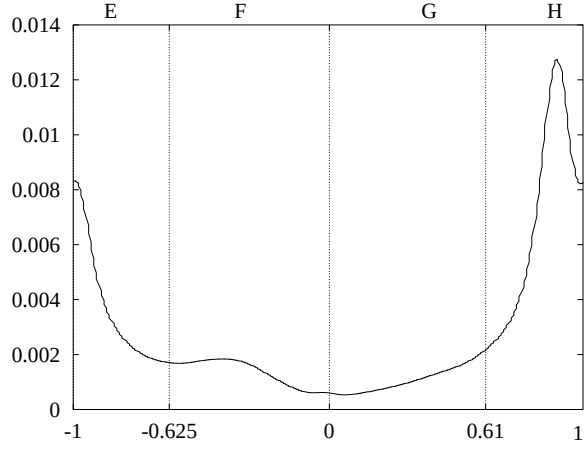


TCO: cosine of carboxyls

- Circular dichroism measurements are mainly sensitive to the cosing of the angle between adjacent backbone carboxyl groups:

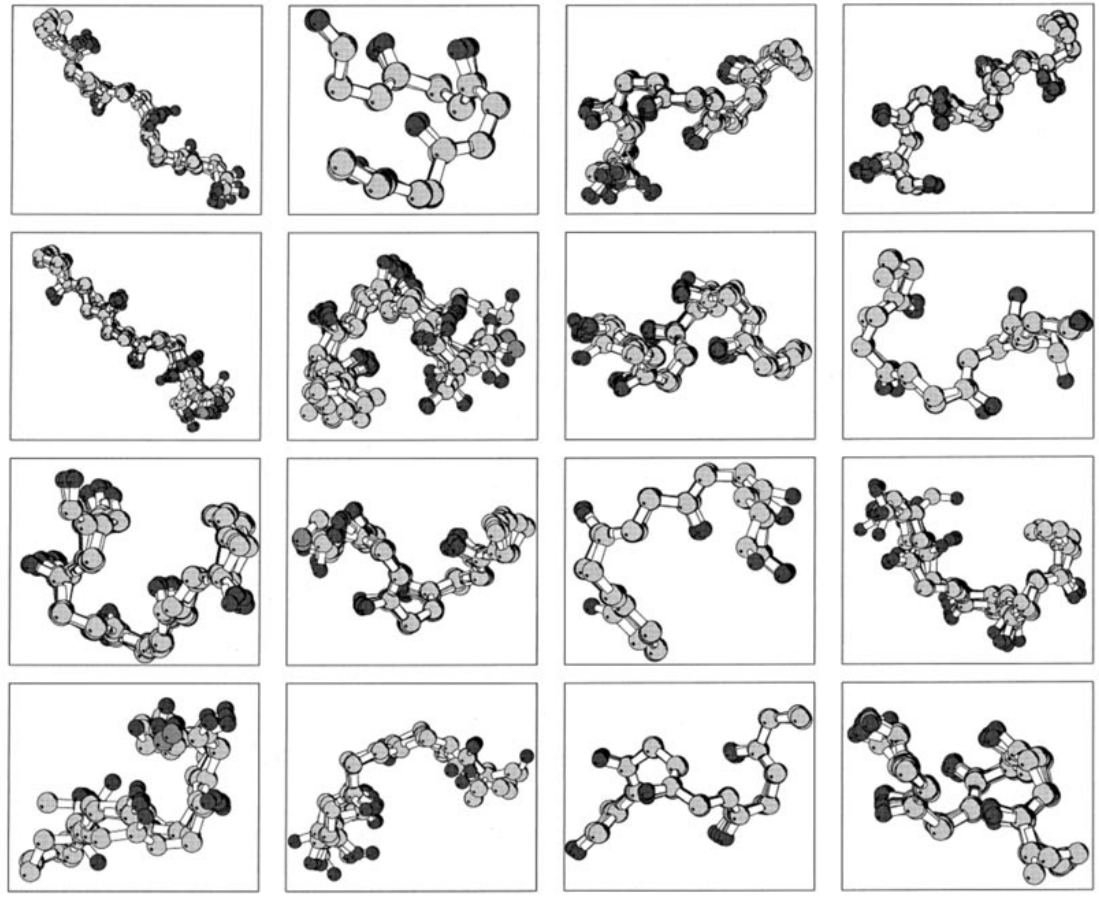


- We used k-means to get 4-letter alphabet:



de Brevern's Protein Blocks

Clustered on 5-residue window of ϕ-ψ angles:



What makes an alphabet good?

A good alphabet should

- capture a conceptually interesting property.
- be assignable by a program.
- be well-conserved during evolution.
- be predictable from amino acid sequence (or profile).
- be useful in improving fold recognition.
- be useful in improving alignment of remote homologs.

Test Sets

We have three sets of data for testing

- A set of multiple alignments based on 3D-structure alignment. (Based on FSSP, Z>=7.0)
- A diverse set of good-quality protein structures, with no more than 30% residue identity, split into 3 sets for 3-fold cross-validation. Taken from Dunbrack's culledPDB lists, further selected to contain domains in SCOP version 1.55.
- A set of difficult pairwise alignment problems, with "correct" alignments determined by several structural aligners.

Protocol

- Make multiple alignment of homologs for each protein (using SAM-T2K or psi-blast).
- Make local-alphabet sequence string for each protein.
- Check conservation using FSSP alignments.
- Train neural nets to predict local structure from SAM-T2K alignment. Measure predictability using 3-fold cross-validation.
- Use SAM-T2K alignment and predicted local structure to build multi-track HMM for each protein and use for all-against-all fold-recognition tests.
- Use the multi-track HMMs to do pairwise alignments and score with shift score.

Conservation check

- FSSP alignments are master-slave alignments.
- We compute mutual information between the local structure label of the master sequence and the local structure labels of the slave sequences in the same alignment column.
- Make a contingency table counting all pairs of labels and compute mutual information of the pairs.
- Mutual information:

$$\mathrm{MI} = \sum_{i,j} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

- We also correct for small sample sizes, but this correction is tiny for small alphabets.

Predictability check

- Neural net output is interpreted as probability vector over local structure alphabet.
- Use neural nets with fixed architecture (4 layers with softmax on each layer, with window sizes of 5,7,9,13 and 15,15,15,|A| units).
- Train on 2/3 of data to maximize $\sum \log P_{NN}(\text{observed letter})$, test on remaining third.
- Compute information gain for test set:

$$\frac{1}{N} \sum \log_2 \frac{P_{NN}(\text{observed letter})}{P_{\emptyset}(\text{observed letter})} \ ,$$

where P_{NN} is the neural net output, $P_{\emptyset}$ is the background probability, and N is the size of the test set.

Predictability (other measures)

- We also look at less interesting measures:
  - $Q_{|A|}$, the fraction of positions correctly predicted (that is, the correct letter has highest probability).
  - SOV, a complicated segment-overlap measure often used in testing EHL predictions.
- $Q_{|A|}$ and SOV are very dependent on the size of the alphabet, making comparison between alphabets difficult.
- Both consider only the letter predicted with highest probability, throwing out all other information in the probability vector.

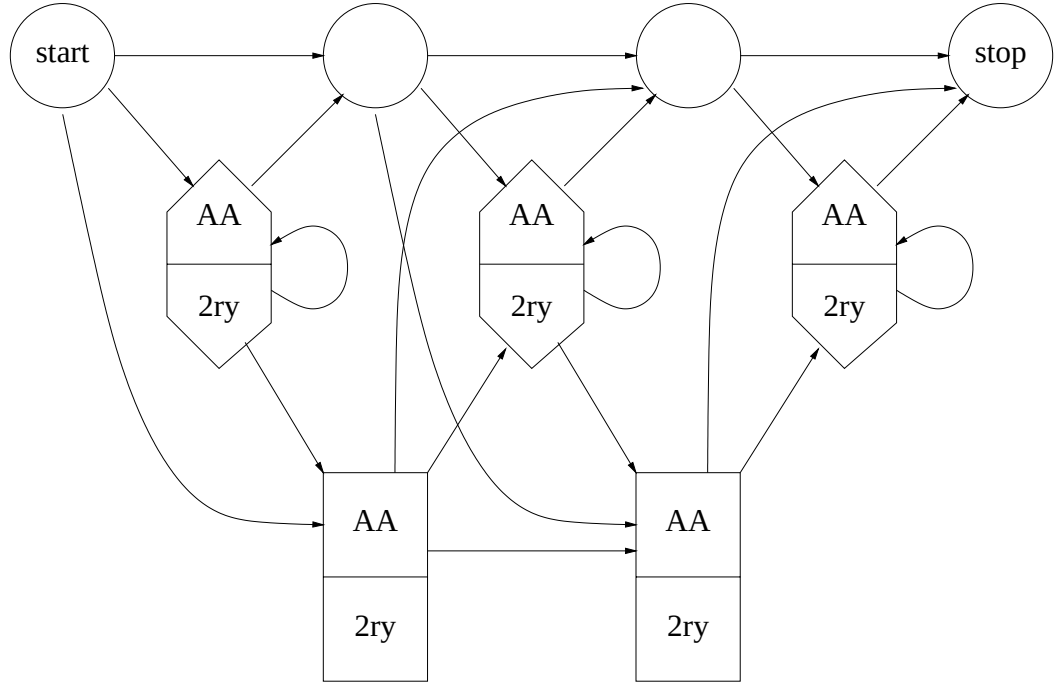
Conservation and Predictability

			conservation	predictability		
Name	alphabet size	entropy	mutual info	info gain per residue	$Q_{\|A\|}$	SOV
str	13	2.842	1.107	1.009	0.561	0.527
protein blocks	16	3.233	0.980	1.259	0.579	0.542
stride	6	2.182	0.904	0.863	0.663	0.659
dssp	7	2.397	0.893	0.913	0.633	0.610
stride-ehl	3	1.546	0.861	0.736	0.769	0.733
dssp-ehl	3	1.545	0.831	0.717	0.763	0.732
alpha11	11	2.965	0.688	0.711	0.469	0.375
Bystroff	10	2.443	0.678	0.736	0.588	0.501
TCO	4	1.810	0.623	0.577	0.649	0.547

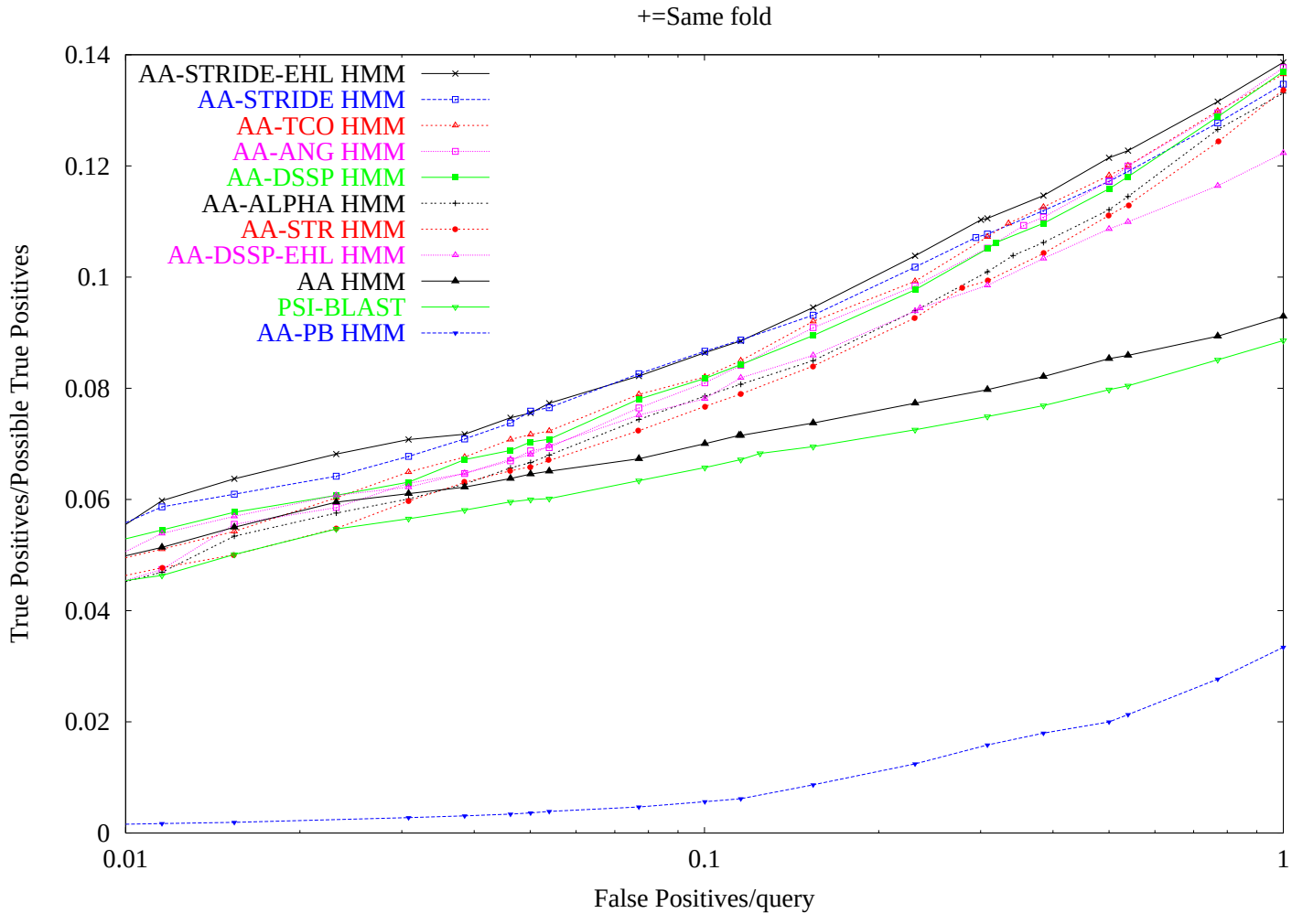
Multi-track HMMs

Use SAM-T2K alignments to build a two-track target HMM:

- Amino-acid track (created from the multiple alignment).
- Local-structure track (probabilities from neural net).
- Score all sequences with all models.



Fold recognition results



Alignment test

- Make two-track HMM for each sequence in alignment pairs.
- Use the HMMs to align the pair of sequences (using posterior-decoded alignment).
- Compare alignments from HMMs to reference alignments from structure-structure aligners.
- Note: have two HMM-based alignments per sequence pair—take the mean of the scores.
- Use two or more different structure-structure aligners to create references.

Shift-score

The shift-score of two alignments X and Y

$$\begin{aligned}
\text{shift_score} &= \frac{\sum_{i=1}^{|X|} cs(X_i)}{|X| + |Y|} \\
\text{where } \epsilon &= \text{small algorithmic parameter, 0.2} \\
|X| &= \text{Number of aligned residue pairs in alignment } X \\
X_i &= \text{Aligned residue pair } i \text{ in alignment } X \\
s(r_i) &= \text{Subscore for residue } r_i \\
&= \begin{cases} \frac{1+\epsilon}{1+|\text{shift}(r_i)|} - \epsilon & \text{if shift}(r_i) \text{ is defined} \\ 0 & \text{otherwise} \end{cases} \\
X_i(A) &= \text{Sequence } A \text{ residue aligned in column } X_i \\
cs(X_i) &= \text{Column score for column } i \text{ in alignment } X \\
&= \begin{cases} s(X_i(A)) + s(X_i(B)) \\ \text{if column } X_i \text{ aligns } X_i(A) \text{ and } X_i(B) \\ \text{0 otherwise} \end{cases}
\end{aligned}$$

Shift score example

Basic depiction of alignment shift

Reference	
template	ABCD--EFG
target	L-MNOPQR-

Candidate	
template	-AB-CDEFG
target	LMNOP--QR

Target Residue	Template residue aligned to in Reference alignment	Template residue aligned to in Candidate alignment	Shift
M	C	A	-2
N	D	B	-2
Q	E	F	+1
R	F	G	+1

Shift score results

	Difficult set		Moderate set	
Reference alignment	Dali	CE	Dali	CE
Dali		0.607		0.616
STR	0.320	**0.307**	**0.466**	**0.418**
Protein blocks	0.309	0.303	0.435	0.395
DSSP	0.306	0.295	0.454	0.402
STRIDE	**0.357**	0.292	0.452	0.400
STRIDE-EHL	0.298	0.290	0.438	0.396
DSSP-EHL	0.297	0.287	0.435	0.391
Alpha11	0.288	0.279	0.429	0.387
Bystroff	0.286	0.276	0.422	0.407
TCO	0.284	0.276	0.421	0.374
one-track amino-acid-only				
SAM-T2K seed	0.220	0.219	0.365	0.325
FSSP seed	0.219	0.192	0.415	0.330

Web sites

UCSC bioinformatics info:

`http://www.soe.ucsc.edu/research/compbio/`

SAM tool suite info:

`http://www.soe.ucsc.edu/research/compbio/sam.html`

HMM **servers:** `http://www.soe.ucsc.edu/research/compbio/HMM-apps/`

These slides:

`http://www.soe.ucsc.edu/~karplus/papers/local-structure-germany02.pdf`