Figures and Tables for Incremental Window-based Protein Sequence Alignment Algorithms

Huzefa Rangwala and George Karypis

Department of Computer Science & Engineering, University of Minnesota,Minneapolis, MN 55455

{rangwala,karypis} @cs.umn.edu

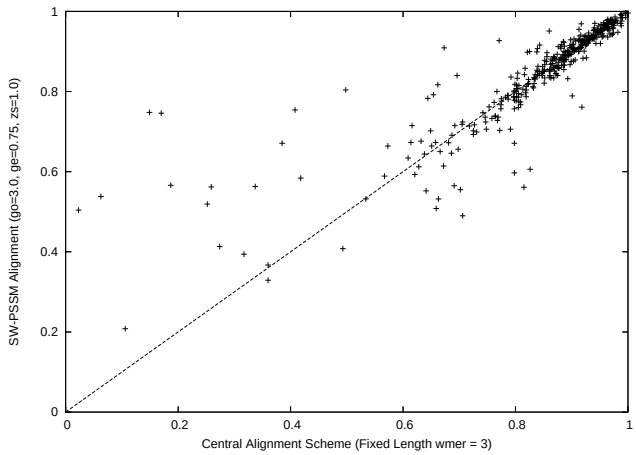

Figure 1: Cline Score Comparison of SW-PSSM scheme against SA^f scheme for the 588 alignment pairs in the template-based dataset

Table 1: Alignment Accuracy Results on a Template-based Dataset.

	f_M (precision)	f_D (recall)	CS	$CS_{\leq 15\%}$
fixed				
central				
$w$mer = 2	0.805	0.791	0.803	0.600
$w$mer = 3	0.799	0.776	0.794	0.596
$w$mer = 4	0.791	0.756	0.782	0.587
$w$mer = 5	0.776	0.732	0.764	0.572
subset				
$w$mer = 2	0.802	0.835	0.826	0.626
$w$mer = 3	0.805	0.842	0.831	0.642
$w$mer = 4	0.805	0.842	0.832	0.644
$w$mer = 5	0.802	0.838	0.828	0.649
combined				
$w$mer = 2	0.791	0.822	0.816	0.619
$w$mer = 3	0.785	0.819	0.814	0.623
$w$mer = 4	0.779	0.811	0.808	0.624
$w$mer = 5	0.767	0.798	0.798	0.624
variable				
central				
$w$mer = 2	0.799	0.804	0.809	0.595
$w$mer = 3	0.802	0.807	0.812	0.605
$w$mer = 4	0.805	0.797	0.810	0.611
$w$mer = 5	0.805	0.797	0.807	0.614
subset				
$w$mer = 2	0.798	0.827	0.820	0.615
$w$mer = 3	0.798	0.834	0.825	0.629
$w$mer = 4	0.798	0.836	0.827	0.634
$w$mer = 5	0.794	0.832	0.823	0.636
combined				
$w$mer = 2	0.795	0.822	0.813	0.600
$w$mer = 3	0.797	0.827	0.820	0.614
$w$mer = 4	0.800	0.831	0.824	0.621
$w$mer = 5	0.800	0.832	0.825	0.628

In the table f_M denotes the Modeler's score, f_D denotes the Developer's score, CS denotes the Cline score, and $CS_{\leq 15\%}$ denotes the Cline score for a subset of sequence pairs sharing less than 15% sequence identity.

Table 2: Alignment Accuracy Results on a Model-based Dataset.

Alignment Scheme	Family	Superfamily	Fold
CA^f (2)	14.86	1.66	0.04
SA^f (5)	16.44	4.29	1.53
CSA^f (2)	15.47	2.53	0.203
CA^v (5)	15.10	2.43	0.12
SA^v (5)	16.48	4.05	1.05
CSA^v (5)	14.05	2.32	0.14

The numbers in the parameter indicate the $w$mer length for the various alignment schemes.

Table 3: Comparative Performance with Earlier Results on Template-based Dataset.

Alignment Scheme	f_M	f_D	CS	$CS_{\leq 15\%}$
SA^f (3)	0.805	0.842	0.831	0.642
SA^v (4)	0.798	0.836	0.827	0.634
SW-PSSM	0.803	0.852	0.841	0.689
pdotp (T99)	0.806	0.829	0.832	0.697
correlp (T99)	0.794	0.835	0.829	0.702
coach (T99)	0.797	0.830	0.829	0.697

The optimized SW-PSSM results are achieved using $go = 3.0$, $ge = 0.75$, $zs = 1.0$. In the table pdotp, correlp, coach use a dot product, correlation function, and a HMM based profile-profile scoring function. T99 denotes the use of SAM T99 based profiles respectively.

Table 4: Comparative Performance with Earlier Results on a Model-based Dataset.

Alignment Scheme	Family	Superfamily	Fold
SA^f (5)	16.44	4.29	1.53
SA^v (5)	16.48	4.05	1.05
SW-PSSM	16.66	4.38	2.02
local	14.1	2.0	0.7
global	15.1	2.9	1.4
PSI	15.8	3.3	1.4
SSPSI	16.0	4.1	2.6
structural	19.4	9.1	8.0

The optimized SW-PSSM results are achieved using $go = 3.0$, $ge = 0.75$, $zs = 3.0$. All the results are optimized for their relevant parameters

Table 5: Fraction of Correct Models based on the LGscore.

LGscore	$< 10^{-3}$			$< 10^{-5}$		
Alignment Scheme	Fm	Sf	Fd	Fm	Sf	Fd
SA^f (3)	74	27	5	55	8	0
SA^v (3)	74	28	4	55	8	0
SW-PSSM	74	27	6	56	8	0
local	66	10	1	46	2	0
global	70	12	1	49	3	0
PSI	72	18	4	50	4	0
SSPSI	73	21	6	53	5	0
structural	86	60	51	66	21	21

The optimized SW-PSSM results are achieved using $go = 3.0$, $ge = 0.75$, $zs = 3.0$. All the results are optimized for their relevant parameters. Fm, Sf and Fd denote the family-level, superfamily-level and fold-level performance results respectively.

Table 6: Reliability Assessment: Recall for the first $k\%$ errors.

Method	0%	1%	2%	3%	4%	5%
CA^f (3)	0.176	0.281	0.365	0.434	0.494	0.541
SA^f (3)	0.186	0.297	0.384	0.459	0.519	0.563
CSA^f (3)	0.180	0.286	0.370	0.438	0.498	0.545
CA^v (3)	0.254	0.364	0.450	0.515	0.566	0.603
SA^v (3)	0.260	0.368	0.454	0.521	0.572	0.612
CSA^v (3)	0.260	0.367	0.454	0.520	0.571	0.610
SW-PSSM	0.205	0.320	0.405	0.480	0.541	0.586

The optimized SW-PSSM results are achieved using $go = 3.0$, $ge = 0.75$, $zs = 3.0$. The numbers in the parenthesis represent the $w$mer width used for the results shown.