Supplementary Data
Reduced amino acid alphabets improve the sensitivity and selectivity of pairwise sequence alignments

Eric L. Peterson
Department of Physics, California Institute of Technology, Pasadena, CA 91125

Jané Kondev
Department of Physics, Brandeis University, Waltham, MA 02454

Julie A. Theriot
Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305

Rob Phillips[*]
Department of Applied Physics, California Institute of Technology, Pasadena, CA 91125

(Dated: September 16, 2008)

Contents

[*] Electronic address: phillips@pboc.caltech.edu

I. ADDITIONAL RESULTS

A. Mean pooled precision

Precision vs. recall curves are shown in panel A of Fig. S1 for GBMR4, HSDM17 and SDM12; the mean pooled precision is the area under this curve. The mean pooled precision for all of the HSDM, SDM, and GBMR alphabets is plotted in panel B of Fig. S1.

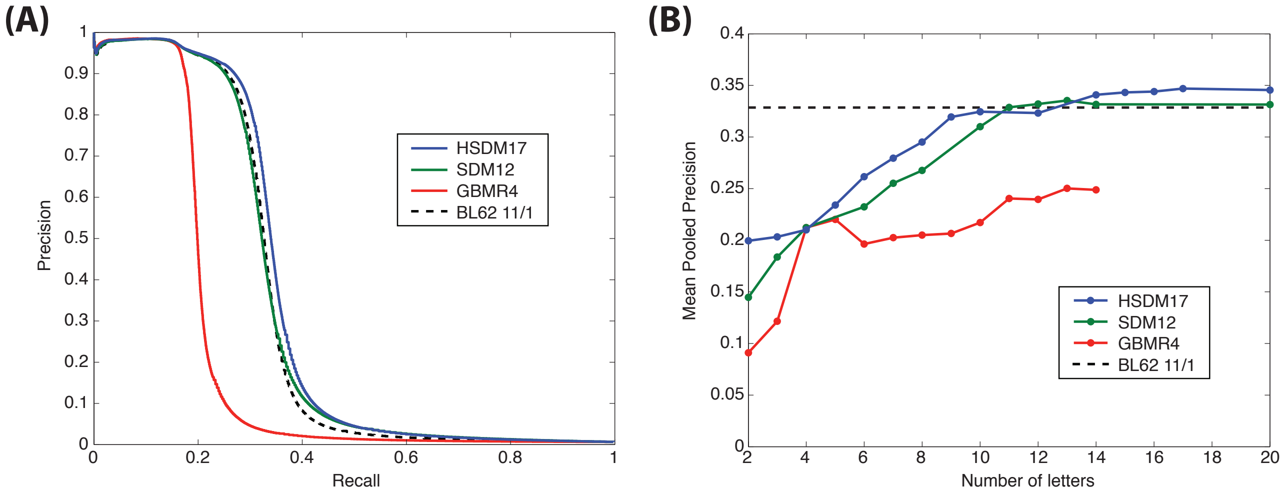

Figure S1: **Reduced alphabet performance in mean pooled precision** (A) Precision vs. recall curves for the top reduced alphabet performers. The mean pooled precision is the area under this curve and indicates the ability of a particular matrix to maintain high selectivity over a wide range of error rates. At some point, each matrix loses the ability to selectively reject false positives and the curve drops precipitously to low precision values. (B) Mean pooled precision indicates the average precision achieved by a matrix over the entire range of recall. Points indicate reduced alphabets that were tested; the connecting lines are a guide to the eye. A perfect method would achieve a mean pooled precision value of unity, with all true positives ranked ahead of false ones. The HSDM17 matrix is the top performer in this metric; the dashed black lines in panels A and B show the performance of BL62 11/1 for reference.

B. Area under the ROC curve

Receiver Operating Characteristic curves are shown in Fig. S2(A) for SDM12, HSDM17 and GBMR4. The total area under the curve vs. number of letters in these schemes is shown in panel B.

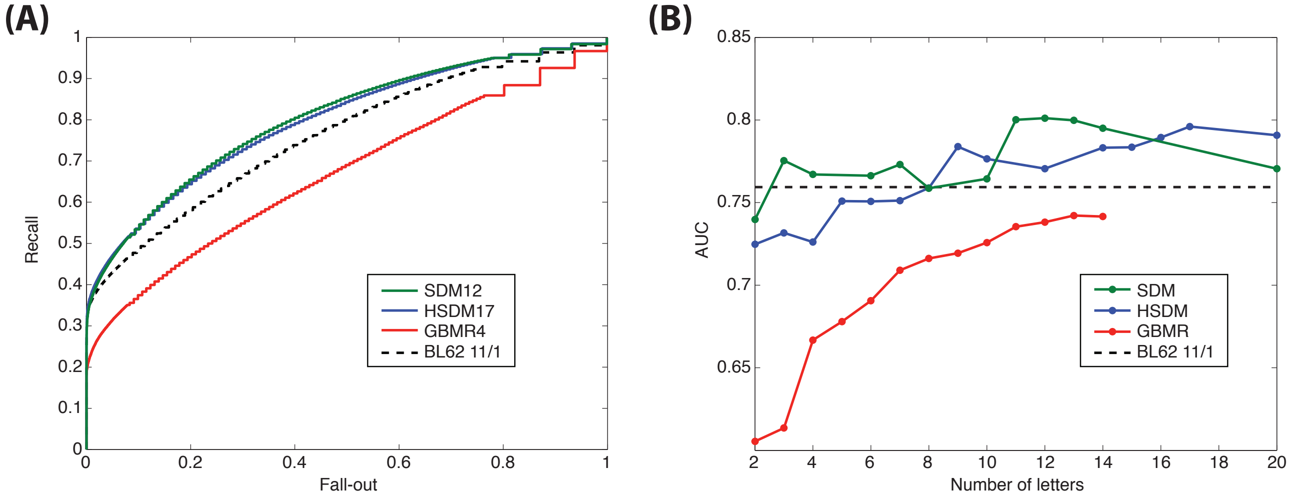

Figure S2: **Reduced alphabet performance in area under the Receiver Operating Characteristic curve.** (A) Receiver Operating Characteristic (ROC) curves for the top performing alphabets. The integral of this curve gives a measure of how well the entire pooled list of hits is sorted; a perfect method would have an ROC area of unity. (B) Overall sensitivity of the SDM alphabets as measured by the area under the ROC curve. The level of sensitivity of BL62 11/1 is shown with the black dashed line. Points indicate reduced alphabets that were tested; the connecting lines are a guide to the eye.

C. Recall at 0.01 EPQ

Panel A of Fig. S3 shows the recall vs. error rate curves under linear normalization for GBMR4, HSDM17 and SDM12 with better-performing matrices generating curves that tend toward the lower-right hand corner indicating high recall at low error rates. Comparing this with panel A of Fig. S1 we can see that GBMR4 is able to maintain the highest level of precision initially, but it rapidly loses precision at higher recall values. Panel B of Fig. S3 shows the recall at 0.01 EPQ with linear normalization vs. number of letters for the GBMR, SDM and HSDM alphabets.

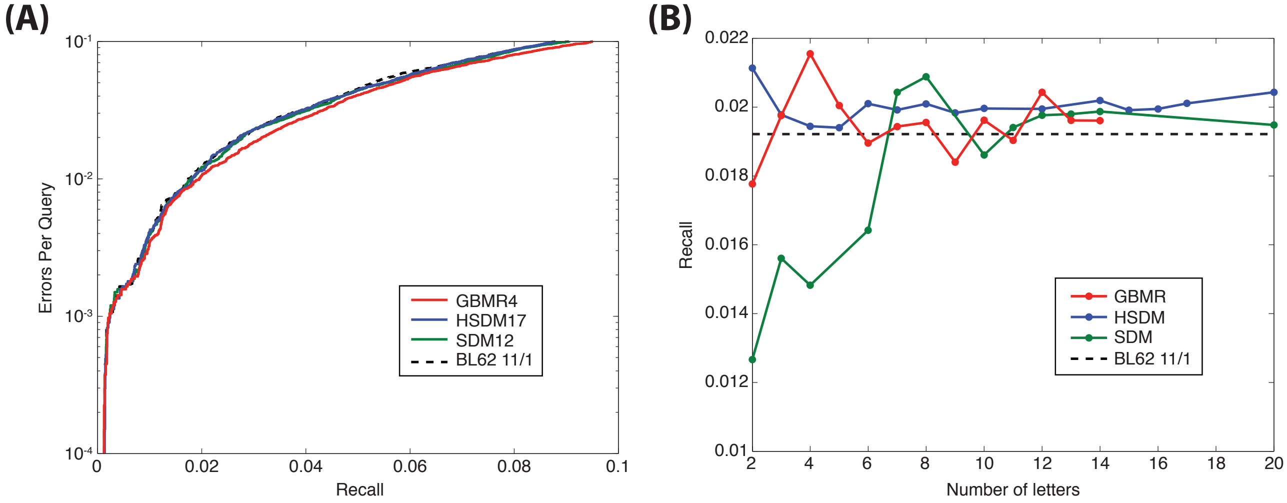

Figure S3: **Reduced alphabet performance in recall vs. errors per query with linear normalization.** (A) Linearly normalized recall (or coverage) vs. the number of errors per query (EPQ). Curves that tend toward the lower right-hand corner perform better, detecting more true positives at a given error rate. Small alphabets show good performance at lower error rates (EPQ < 0.1) with GBMR4 being the top performer. (B) Recall with linear normalization at 0.01 EPQ for various numbers of letters in the GBMR, HSDM and SDM reduced alphabet schemes. The level of performance of BL62 11/1 is shown with the black dashed line. Points indicate reduced alphabets that were tested; the connecting lines are a guide to the eye.

II. ALIGNMENT ACCURACY

We evaluated how well pairwise sequence alignments with reduced alphabets identified pairs of residues that are structurally equivalent as defined by DALI. The results are shown in Fig. S4, plotted as the fraction of structurally equivalent residue pairs identified by SSEARCH using the SDM, HSDM and GBMR reduced alphabet schemes. The curves tend to saturate at around 10 letters, implying that expanding the alphabet beyond this point does not improve the alignments but tends to increase their sensitivity to more recently diverged proteins. The top 10 finishers in alignment accuracy are shown in Table I; HSDM and SDM show the best performance which is not surprising given that they were derived from structurally equivalent pairs of residues [2]. It is interesting that the highly simplified GBMR4 alphabet is able to achieve nearly the same level of accuracy as the full BL62 11/1 matrix. The DALI database of structurally equivalent residues is an exceedingly challenging test of pairwise sequence comparison since the equivalenced residues share only 11% identity overall; even the best alphabet, HSDM17, achieves exact agreement with less than one tenth of all residues in the DALI structural alignments.

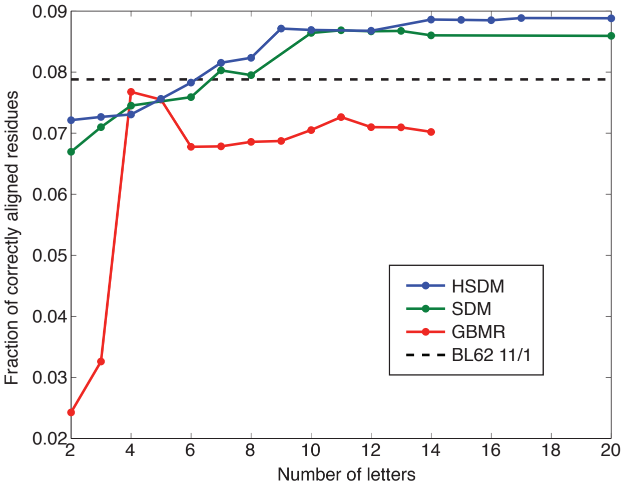

Figure S4: Agreement of structural and sequence alignments. The fraction of DALI equivalent residue pairs found by SSEARCH alignment is shown for various reduced alphabet schemes. Most of the gains are made as classes are added up until around 10 classes, after which the performance levels off. Points indicate reduced alphabets that were tested; the connecting lines are a guide to the eye.

Rank	Scheme	Letters	Fraction aligned
1	HSDM	17	0.08887
2	HSDM	20	0.08882
3	HSDM	14	0.08862
4	HSDM	15	0.08857
5	HSDM	16	0.08849
6	HSDM	9	0.08714
7	HSDM	10	0.08691
8	SDM	11	0.08686
9	HSDM	12	0.08676
10	SDM	13	0.08675

TABLE I: The top 10 performers in agreement between sequence and structural alignments, using DALI structurally equivalent residues as the “gold standard”. As expected, the two structure-derived matrices, HSDM and SDM, completely dominate the results.

III. COMPARISON OF DETECTED RELATIONSHIPS

It is also valuable to compare the hits returned by two matrices at a given errors per query level to see what types of relationships are more easily detected by one relative to another. We compared the hits returned by the SDM12 and BL62 11/1 matrices at or above 0.01 EPQ and found that each matrix finds about 3000 true positives at that error level. After separating out the hits that were unique to each matrix (they share 2724 hits in common) SDM12 was left with 271 unique hits and BL62 11/1 with 139. The approximate mean percent identity of the SDM12 unique hits is 60% whereas for BL62 11/1 it is 70%. Although SDM12 and BL62 have essentially identical relative entropy (-0.703 and -0.699 bits respectively) SDM12 is able to detect more distant relationships than BL62. A histogram of the hits unique to each matrix is shown in Fig. S5.

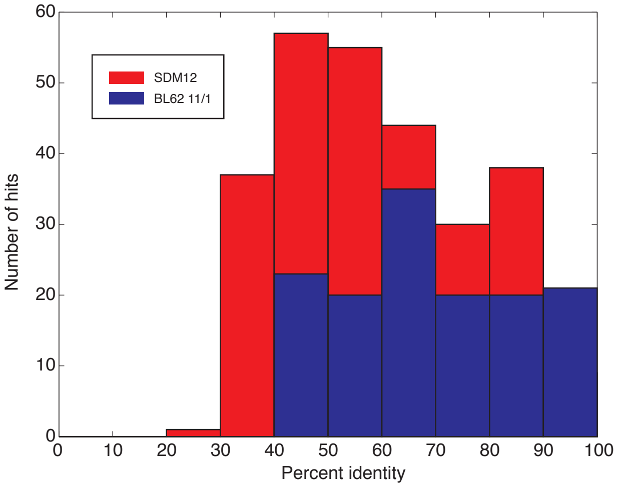

Figure S5: Histogram of the hits at or above 0.01 errors per query unique to the SDM12 and BL62 11/1 matrices. The results from SDM12 are both more numerous and shifted towards lower identity, showing its increased ability to detect more remote relationships.

IV. PRELIMINARY SCOP STUDY RESULTS

In the main text we refer to a study of all vs. all alignments with HSDM17, SDM12, GBMR4 and BL62 11/1 using proteins belonging to the same SCOP superfamily to define true positives [1]. The complete results of this small-scale study are shown in Table II below, with best results in bold. The HSDM17 and SDM12 schemes maintain an advantage over BL62 11/1 in the SCOP study, while the smaller GBMR4 alphabet does not perform as well relative to BL62 11/1 as it did in the DALI study. At both the 40% and 95% levels of identity, the HSDM17 alphabet performed best with the SCOP databases in mean precision, area under the ROC curve and linearly normalized recall. The SCOP database includes curation by experts to make judgements about the evolutionary relationships between proteins, whereas DALI uses structural similarity alone as the criteria for determining relatedness.

	scop40			scop95			DALI		
Scheme	MPP	AUC	Recall	MPP	AUC	Recall	MPP	AUC	Recall
GBMR4	0.089	0.658	0.100	0.259	0.727	0.204	0.212	0.667	**0.022**
SDM12	0.156	0.734	0.136	0.419	0.833	0.250	0.332	**0.801**	0.020
HSDM17	**0.173**	**0.751**	**0.148**	**0.436**	**0.840**	**0.259**	**0.347**	0.796	0.020
BL62 11/1	0.156	0.714	0.134	0.408	0.812	0.245	0.329	0.759	0.019

TABLE II: Comparison of results from all vs. all studies with scop40, scop95 and DALI. In the SCOP results GMBR4 is unable to maintain its advantage in linearly normalized recall at 0.01 EPQ over BL62 11/1. However both SDM12 and HSDM17 are able to match or better the results of BL62 11/1 in mean pooled precision (MPP), area under the ROC curve (AUC) and linearly normalized recall at 0.01 EPQ. Version 1.71 of the scop40 and scop95 sequence databases were used.

V. DALI STUDY RESULTS

A. Top 10 performers

	(A) Mean pooled precision			(B) Area under ROC curve			(C) Recall at 0.01 EPQ		
Rank	Scheme	Groups	MPP	Scheme	Groups	AUC	Scheme	Groups	Recall
1	HSDM	17	0.347	SDM	12	0.801	GBMR	4	0.022
2	BL62 7/1	20	0.347	SDM	11	0.800	CB/LW	2	0.021
3	BL50 11/1	20	0.346	SDM	13	0.800	HSDM	2	0.021
4	LZ-BL	16	0.346	HSDM	17	0.796	LZ-BL	7	0.021
5	HSDM	20	0.345	SDM	14	0.795	SDM	8	0.021
6	HSDM	16	0.344	LZ-MJ	6	0.793	TD	2	0.021
7	HSDM	15	0.343	HSDM	20	0.791	BL50 11/1	20	0.021
8	HSDM	14	0.341	HSDM	16	0.789	LZ-BL	6	0.020
9	LZ-BL	15	0.339	HSDM	9	0.784	HSDM	20	0.020
10	SDM	13	0.335	HSDM	15	0.783	SDM	7	0.020

TABLE III: Top 10 performers in mean pooled precision (MPP), area under the Receiver Operating Characteristic curve (AUC) and recall at 0.01 errors per query with linear normalization. Mean pooled precision is a measure of the selectivity of a matrix i.e. its ability to retain high recall of true positive relationships at low error rates. The area under the ROC curve measures the sensitivity of a matrix to true positive alignments over the entire list of results. Recall at 0.01 EPQ measures the selectivity of a matrix but is drawn from a limited set of hits such as a researcher might reasonably peruse manually.

B. Complete DALI study results

TABLE IV: Results for all alphabets and matrices tested

					Recall at 0.01 EPQ		
Scheme	**Letters**	**AUC**	**MPP**	**Align**	**None**	**Linear**	**Quadratic**
AB	2	0.665	0.159	0.04844	0.0026	0.020	0.034
AB	3	0.674	0.178	0.04968	0.0025	0.019	0.033
AB	4	0.682	0.194	0.05828	0.0026	0.020	0.034
AB	5	0.702	0.206	0.06143	0.0025	0.019	0.033
AB	6	0.719	0.210	0.05877	0.0024	0.019	0.034
AB	7	0.731	0.234	0.06827	0.0025	0.019	0.033
AB	8	0.721	0.252	0.07310	0.0026	0.020	0.034
AB	9	0.728	0.294	0.07554	0.0026	0.019	0.033
AB	10	0.740	0.297	0.07509	0.0026	0.020	0.034
AB	11	0.749	0.310	0.07690	0.0027	0.020	0.034
AB	12	0.749	0.313	0.07736	0.0027	0.020	0.034
AB	13	0.750	0.312	0.07610	0.0027	0.020	0.034
AB	14	0.755	0.314	0.07599	0.0026	0.020	0.034
AB	15	0.753	0.319	0.07603	0.0026	0.020	0.033
AB	16	0.754	0.320	0.07648	0.0026	0.019	0.033
AB	17	0.752	0.322	0.07702	0.0026	0.019	0.033
AB	18	0.756	0.326	0.07876	0.0026	0.020	0.034
AB	19	0.757	0.323	0.07828	0.0026	0.019	0.033
BL50 11/1	20	0.779	0.346	0.08476	0.0027	0.021	0.035
BL50 12/2	20	0.762	0.334	0.08111	0.0026	0.020	0.034
BL62 11/1	20	0.759	0.329	0.08322	0.0025	0.019	0.033
CB	2	0.705	0.188	0.06774	0.0029	0.021	0.035
CB	5	0.674	0.191	0.05904	0.0024	0.018	0.031
DSSP	2	0.679	0.154	0.05393	0.0044	0.019	0.033
DSSP	3	0.731	0.209	0.07479	0.0025	0.020	0.034
DSSP	4	0.709	0.222	0.07996	0.0027	0.020	0.033
DSSP	5	0.723	0.219	0.07736	0.0026	0.019	0.032
DSSP	6	0.729	0.230	0.07913	0.0025	0.019	0.033
DSSP	7	0.738	0.246	0.08042	0.0025	0.019	0.032
DSSP	8	0.730	0.233	0.07572	0.0024	0.018	0.031
DSSP	9	0.731	0.244	0.07631	0.0024	0.019	0.033
DSSP	10	0.733	0.253	0.07740	0.0025	0.019	0.032
DSSP	11	0.757	0.282	0.07829	0.0026	0.019	0.033
DSSP	12	0.759	0.287	0.07942	0.0026	0.019	0.033
DSSP	13	0.758	0.290	0.08084	0.0026	0.019	0.033
DSSP	14	0.768	0.297	0.08252	0.0026	0.020	0.034
GBMR	2	0.605	0.091	0.02423	0.0029	0.018	0.032
GBMR	3	0.614	0.122	0.03261	0.0025	0.020	0.035

TABLE IV: Results for all alphabets and matrices tested (continued)

					Recall at 0.01 EPQ		
Scheme	**Letters**	**AUC**	**MPP**	**Align**	**None**	**Linear**	**Quadratic**
GBMR	4	0.667	0.212	0.07676	0.0029	0.022	0.036
GBMR	5	0.678	0.220	0.07549	0.0027	0.020	0.033
GBMR	6	0.691	0.196	0.06778	0.0097	0.019	0.031
GBMR	7	0.709	0.202	0.06784	0.0089	0.019	0.032
GBMR	8	0.716	0.205	0.06857	0.0085	0.020	0.032
GBMR	9	0.719	0.206	0.06871	0.0067	0.018	0.030
GBMR	10	0.726	0.217	0.07052	0.0038	0.020	0.033
GBMR	11	0.735	0.240	0.07264	0.0028	0.019	0.032
GBMR	12	0.738	0.240	0.07098	0.0027	0.020	0.035
GBMR	13	0.742	0.250	0.07096	0.0026	0.020	0.034
GBMR	14	0.742	0.249	0.07022	0.0026	0.020	0.034
HSDM	2	0.725	0.199	0.07214	0.0029	0.021	0.036
HSDM	3	0.732	0.203	0.07266	0.0026	0.020	0.034
HSDM	4	0.726	0.210	0.07306	0.0026	0.019	0.034
HSDM	5	0.751	0.234	0.07562	0.0027	0.019	0.034
HSDM	6	0.751	0.262	0.07827	0.0027	0.020	0.035
HSDM	7	0.751	0.279	0.08154	0.0028	0.020	0.033
HSDM	8	0.759	0.295	0.08235	0.0027	0.020	0.034
HSDM	9	0.784	0.319	0.08714	0.0028	0.020	0.033
HSDM	10	0.776	0.325	0.08691	0.0027	0.020	0.034
HSDM	12	0.771	0.323	0.08676	0.0026	0.020	0.035
HSDM	14	0.783	0.341	0.08862	0.0027	0.020	0.035
HSDM	15	0.783	0.343	0.08857	0.0026	0.020	0.035
HSDM	16	0.789	0.344	0.08849	0.0026	0.020	0.035
HSDM	17	0.796	0.347	0.08887	0.0027	0.020	0.035
HSDM	20	0.791	0.345	0.08882	0.0026	0.020	0.036
JO20	20	0.725	0.274	0.05900	0.0024	0.019	0.033
LR	10	0.719	0.280	0.07019	0.0026	0.020	0.034
LW-I	2	0.705	0.188	0.06774	0.0029	0.021	0.035
LW-I	3	0.720	0.203	0.06403	0.0027	0.020	0.034
LW-I	4	0.697	0.232	0.07038	0.0026	0.019	0.032
LW-I	5	0.701	0.227	0.06388	0.0027	0.020	0.034
LW-I	6	0.695	0.241	0.06369	0.0027	0.020	0.034
LW-I	7	0.688	0.240	0.06681	0.0026	0.020	0.033
LW-I	8	0.737	0.286	0.07320	0.0026	0.020	0.034
LW-I	9	0.740	0.290	0.07389	0.0026	0.020	0.035
LW-I	10	0.728	0.292	0.07218	0.0025	0.019	0.033
LW-I	11	0.735	0.303	0.07579	0.0026	0.020	0.035
LW-I	12	0.740	0.303	0.07605	0.0026	0.020	0.034
LW-I	13	0.754	0.310	0.07662	0.0026	0.020	0.034

TABLE IV: Results for all alphabets and matrices tested (continued)

					Recall at 0.01 EPQ		
Scheme	**Letters**	**AUC**	**MPP**	**Align**	**None**	**Linear**	**Quadratic**
LW-I	14	0.756	0.307	0.07569	0.0025	0.020	0.034
LW-I	15	0.757	0.308	0.07593	0.0025	0.020	0.034
LW-I	16	0.754	0.314	0.07599	0.0026	0.020	0.034
LW-I	17	0.752	0.317	0.07619	0.0026	0.020	0.034
LW-I	18	0.753	0.318	0.07663	0.0026	0.020	0.034
LW-I	19	0.755	0.321	0.07693	0.0026	0.020	0.034
LW-NI	2	0.705	0.188	0.06774	0.0029	0.021	0.035
LW-NI	3	0.720	0.203	0.06403	0.0027	0.020	0.034
LW-NI	4	0.723	0.224	0.06361	0.0025	0.019	0.032
LW-NI	5	0.702	0.229	0.06417	0.0026	0.019	0.032
LW-NI	6	0.707	0.243	0.06453	0.0026	0.019	0.033
LW-NI	7	0.698	0.245	0.06795	0.0026	0.020	0.033
LW-NI	8	0.698	0.244	0.06509	0.0025	0.020	0.034
LW-NI	9	0.696	0.249	0.06569	0.0025	0.019	0.034
LW-NI	10	0.706	0.263	0.06816	0.0025	0.020	0.034
LW-NI	11	0.739	0.292	0.07406	0.0025	0.020	0.034
LW-NI	12	0.740	0.303	0.07605	0.0026	0.020	0.034
LW-NI	13	0.754	0.310	0.07662	0.0026	0.020	0.034
LW-NI	14	0.756	0.312	0.07688	0.0027	0.020	0.034
LW-NI	15	0.757	0.308	0.07593	0.0025	0.020	0.034
LW-NI	16	0.754	0.314	0.07599	0.0026	0.020	0.034
LW-NI	17	0.752	0.317	0.07619	0.0026	0.020	0.034
LW-NI	18	0.753	0.318	0.07663	0.0026	0.020	0.034
LW-NI	19	0.755	0.321	0.07693	0.0026	0.020	0.034
LZ-BL	2	0.690	0.194	0.06969	0.0026	0.020	0.033
LZ-BL	3	0.719	0.217	0.07751	0.0026	0.019	0.033
LZ-BL	4	0.736	0.242	0.08056	0.0027	0.020	0.033
LZ-BL	5	0.734	0.282	0.08134	0.0028	0.020	0.034
LZ-BL	6	0.742	0.299	0.08337	0.0028	0.020	0.035
LZ-BL	7	0.744	0.297	0.08129	0.0027	0.021	0.036
LZ-BL	8	0.736	0.299	0.08115	0.0026	0.020	0.036
LZ-BL	9	0.744	0.300	0.08092	0.0026	0.020	0.034
LZ-BL	10	0.749	0.327	0.08417	0.0026	0.020	0.034
LZ-BL	11	0.750	0.325	0.08184	0.0026	0.020	0.035
LZ-BL	12	0.769	0.328	0.08326	0.0025	0.019	0.033
LZ-BL	13	0.774	0.331	0.08380	0.0026	0.020	0.034
LZ-BL	14	0.774	0.334	0.08391	0.0025	0.019	0.033
LZ-BL	15	0.777	0.339	0.08413	0.0026	0.020	0.034
LZ-BL	16	0.783	0.346	0.08451	0.0027	0.020	0.034
LZ-MJ	2	0.700	0.165	0.05816	0.0026	0.019	0.033

TABLE IV: Results for all alphabets and matrices tested (continued)

					Recall at 0.01 EPQ		
Scheme	**Letters**	**AUC**	**MPP**	**Align**	**None**	**Linear**	**Quadratic**
LZ-MJ	3	0.666	0.173	0.05389	0.0023	0.018	0.032
LZ-MJ	4	0.722	0.203	0.06992	0.0024	0.019	0.032
LZ-MJ	5	0.779	0.221	0.07165	0.0022	0.017	0.031
LZ-MJ	6	0.793	0.220	0.07124	0.0022	0.018	0.031
LZ-MJ	7	0.770	0.246	0.07434	0.0023	0.018	0.030
LZ-MJ	8	0.750	0.250	0.07749	0.0025	0.019	0.032
LZ-MJ	9	0.757	0.261	0.08010	0.0024	0.018	0.031
LZ-MJ	10	0.764	0.266	0.08065	0.0024	0.018	0.031
LZ-MJ	11	0.759	0.265	0.07871	0.0023	0.018	0.030
LZ-MJ	12	0.782	0.279	0.07966	0.0023	0.018	0.031
LZ-MJ	13	0.782	0.285	0.08017	0.0024	0.018	0.031
LZ-MJ	14	0.783	0.285	0.08082	0.0023	0.018	0.031
LZ-MJ	15	0.773	0.311	0.08207	0.0024	0.019	0.032
LZ-MJ	16	0.773	0.312	0.08259	0.0024	0.019	0.032
ML	4	0.693	0.236	0.07146	0.0026	0.019	0.032
ML	8	0.753	0.294	0.07733	0.0027	0.020	0.035
ML	10	0.757	0.304	0.08087	0.0027	0.020	0.033
ML	15	0.762	0.331	0.08063	0.0027	0.020	0.034
MM	5	0.691	0.210	0.06894	0.0023	0.017	0.030
MS	6	0.715	0.232	0.06853	0.0027	0.020	0.035
SDM	2	0.740	0.145	0.06695	0.0022	0.013	0.020
SDM	3	0.775	0.184	0.07099	0.0022	0.016	0.026
SDM	4	0.767	0.212	0.07450	0.0026	0.015	0.024
SDM	6	0.766	0.232	0.07591	0.0030	0.016	0.029
SDM	7	0.773	0.255	0.08029	0.0028	0.020	0.034
SDM	8	0.759	0.268	0.07952	0.0028	0.021	0.035
SDM	10	0.764	0.310	0.08642	0.0026	0.019	0.031
SDM	11	0.800	0.329	0.08686	0.0027	0.019	0.033
SDM	12	0.801	0.332	0.08670	0.0026	0.020	0.034
SDM	13	0.800	0.335	0.08675	0.0027	0.020	0.034
SDM	14	0.795	0.332	0.08603	0.0027	0.020	0.034
SDM	20	0.770	0.331	0.08594	0.0026	0.019	0.033
TD	2	0.678	0.162	0.05673	0.0027	0.021	0.035
TD	3	0.679	0.162	0.05631	0.0025	0.020	0.034
TD	4	0.704	0.175	0.06090	0.0024	0.019	0.031
TD	5	0.718	0.185	0.06316	0.0023	0.018	0.032
TD	6	0.768	0.224	0.07772	0.0023	0.018	0.031
TD	7	0.740	0.237	0.08096	0.0023	0.018	0.031
TD	8	0.748	0.265	0.08159	0.0024	0.018	0.031
TD	9	0.737	0.278	0.08153	0.0025	0.019	0.032

TABLE IV: Results for all alphabets and matrices tested (continued)

					Recall at 0.01 EPQ		
Scheme	**Letters**	**AUC**	**MPP**	**Align**	**None**	**Linear**	**Quadratic**
TD	10	0.743	0.283	0.08033	0.0025	0.019	0.032
TD	14	0.743	0.306	0.07984	0.0026	0.020	0.034
WW	5	0.709	0.219	0.07172	0.0026	0.019	0.033

[1] 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*, 247(4):536–540, 1995.

[2] A Prlić, F S Domingues, and M J Sippl. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng*, 13(8):545–550, 2000.