Protein Sequence Alignment and Structural Disorder: A Substitution Matrix for an Extended Alphabet

Uros Midic
Information Science and Technology Center, Temple University, Philadelphia, PA
uros@ist.temple.edu

A. Keith Dunker
Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN
kedunker@iupui.edu

Zoran Obradovic[1]
Information Science and Technology Center, Temple University, Philadelphia, PA
zoran@ist.temple.edu

ABSTRACT

In protein sequence alignment algorithms, a substitution matrix of 20x20 alignment parameters is used to describe the rates of amino acid substitutions over time. Development and evaluation of most substitution matrices including the BLOSUM family [1] was based almost entirely on fully structured proteins. Structurally disordered proteins (i.e. proteins that lack structure, either in part or as a whole) that have been shown to be very common in nature [2] have a significantly different amino acid composition than ordered (i.e. structured) proteins [3]. Furthermore, the sequence evolution rate is higher in unstructured as compared to structured regions of proteins containing both structured and unstructured regions [4]. These results cast doubt on appropriateness of the BLOSUM substitution matrices for alignment of structurally disordered proteins [5].To address this problem, we take into the account the concept of structural disorder by extending the alphabet for sequence representation from 20 to 2x20=40 symbols, 20 for amino acids in disordered regions and 20 for amino acids in ordered regions. A 40x40 substitution matrix is required for alignment of sequences represented in the extended alphabet. Such an expanded matrix contains 20x20 submatrices that correspond to matching ordered-ordered, ordered-disordered, and disordered-disordered pairs of residues. In this paper we describe an iterative procedure that we used to estimate such a 40x40 substitution matrix. The iterative procedure converged with stable results with respect to the choice of the sequences in the dataset. In the obtained 40x40 matrix we found substantial differences between the 20x20 submatrices corresponding to ordered-ordered, ordered-disordered, and disordered-disordered region matching. These differences provide evidence that for alignment of protein sequences that contain disordered segments, the discovered substitution matrix is more appropriate than the BLOSUM substitution matrices. At the same time, the new substitution matrix is applicable for sequence alignment of fully ordered proteins as its order-order submatrix is very similar to a BLOSUM matrix.

StReBio'09, June 28, 2009, Paris, France.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences – *Biology and genetics.*

General Terms

Algorithms, Experimentation.

Keywords

Protein sequence alignment; structurally disordered proteins; substitution matrices.

1. INTRODUCTION

Sequence alignment is an essential tool in modern bioinformatics. The goal of sequence alignment is to arrange two or more sequences (genomic or protein) in rows of equal length in an attempt to identify similar and evolutionary related sequences. The alignment process allows mismatching and gaps where mismatches correspond to point mutations while gaps correspond to insertions and deletions. Most alignment algorithms, including BLAST [6] and ClustalW [7], use a matrix of parameters known as *substitution* or *scoring matrix* to assign scores to possible alignments and then look for an alignment with maximal score. Additionally, penalties for gaps can also be controlled with parameters, such as gap opening penalty and gap extension penalty.

Substitution matrices are derived from a set of "ground-truth" alignments; PAM matrices [8] were developed from a set of manually curated alignments, while BLOSUM matrices [1] were developed from alignments in the BLOCKS database [9]. There is no natural golden standard for the choice of set of "ground-truth" alignments, and this choice is one of the main sources of variation between various substitution matrices. The score $score(a_i , a_j)$ for matching amino acids a_i and a_j is calculated as $C \cdot \log_2(p_{ij} / q_i q_j)$, where p_{ij} is the observed frequency of a_i and a_j being aligned in the "ground-truth" alignments, while q_i and q_j are the observed frequencies of a_i and a_j, and the constant C is selected so that the error introduced by rounding all scores to the nearest integer is minimized. The score is positive if amino acids a_i and a_j are

[1] Please send all correspondence to: Zoran Obradovic, Temple University, 303 Wachman Hall (038-24), 1805 N. Broad St., Philadelphia, PA 19122, USA. Phone: +1 (215) 204-6265, Fax: +1 (215) 204-5082, E-mail: zoran@ist.temple.edu

observed aligned as a pair more frequently than would be expected based on their individual frequencies, and negative if they are observed aligned less frequently than would be expected.

Structurally disordered proteins (SDPs, also called *intrinsically disordered proteins/IDPs* or *unstructured proteins*) are highly abundant in nature [2]. Although they lack stable tertiary structure under physiological conditions in vitro, the functional repertoire of SDPs complements the functions of ordered proteins. SDPs are involved in a number of crucial biological functions including regulation, recognition, signaling and control [10]. The structurally disordered regions (SDRs, also called *intrinsically disordered regions/IDRs* or *unstructured regions*) in proteins have significantly different amino acid composition than ordered proteins [3]. This observation led to development of predictors of structural disorder that achieve more than 80% of per-residue accuracy [11]. The difference in amino acid compositions alone casts doubt on appropriateness of BLOSUM and PAM matrices for alignment of SDP sequences (since frequencies q_i are different). Rates of sequence evolution in disordered versus ordered proteins were examined in [4], where it was found that for 19 out of 26 families of proteins with confirmed structural disorder, the disordered regions evolved significantly more rapidly than the ordered regions, while for only 2 families the opposite was true. A different rate of evolution in disordered proteins means that the frequencies p_{ij} are also inappropriate, and a different substitution matrix is needed for alignment of SDP sequences.

To overcome the lack of "ground-truth" alignments for SDPs, an iterative approach has previously been used [5] to obtain a set of alignments of families of proteins with confirmed SDRs and the corresponding substitution matrix. The iterative procedure starts with the BLOSUM62 matrix, aligns all families of proteins and calculates the substitution matrix from obtained alignments. The two steps of alignment and calculation of the substitution matrix are then repeated until no significant changes are observed. The obtained matrix DISORDER is significantly different than the initial BLOSUM62 matrix. However, there is no clear-cut criterion for when this matrix should be used instead of the BLOSUM62 matrix. Furthermore, this matrix still assigns the same score to a pair of amino acids, regardless of whether they belong to SDRs or ordered regions of proteins.

In this paper we propose a radically new approach to protein sequence representation for the purpose of sequence alignment that takes into account the concept of structural disorder and the differences in amino acid compositions and evolutionary rates. We use an extended amino acid alphabet that assigns two different symbols to the same amino acid depending on whether it belongs to a structurally disordered region or a structured region. We describe an iterative procedure that we used to obtain a 40x40 substitution matrix. This matrix has four 20x20 submatrices that correspond to aligning: 1) ordered to ordered regions, 2) and 2') ordered to disordered regions, and 3) disordered to disordered regions (Figure 1). We found significant and substantial differences between these submatrices. The scores for alignment of disordered regions to disordered regions are higher than for alignment of ordered regions to ordered regions, which is further empirical evidence of higher evolutionary rate in disordered regions. The most important advantage of this approach is that the alignment algorithms such as Needleman-Wunsch [12], Smith-Waterman [13] and ClustalW [7] can be modified to use the expanded substitution matrix and utilize the knowledge (experimentally determined or predicted) of structurally disordered regions in the sequences being aligned.

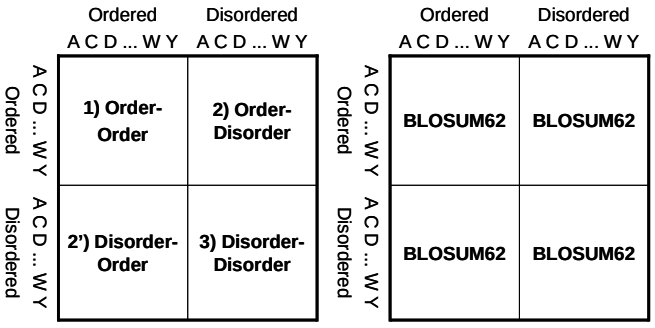

Figure 1. A 40x40 substitution matrix consists of four 20x20 sub-matrices (left) and the initial matrix, made up of four copies of BLOSUM62 matrix (right).

2. MATERIALS AND METHODS

2.1 Dataset

To overcome the limitation on the size of dataset from [5], where only proteins with confirmed SDRs were used, we decided to use prediction of structural disorder to label the SDRs in protein sequences, which in turn allows us to select arbitrary families of protein sequences for our dataset. We began by randomly selecting 1000 protein sequences from the UNIREF database as "anchors" for families. We performed BLAST queries for these sequences against the UNIREF database to obtain families of similar sequences. From the BLAST results we kept only those sequences that satisfied the following criteria: 1) the difference in sequence length compared to the anchor sequence was less than 10%, and 2) the global sequence identity with the anchor sequence was at least 90% (note that significance of BLAST results is estimated based on local identity and/or similarity). We discarded the families with less than 10 sequences. To limit the computational requirements we imposed a threshold of 900 on the length of sequences and reduced the large families to only 50 sequences by random sampling. The resulting dataset contains 600 families with between 10 and 50 sequences (436 families, or 72%, contain 50 sequences). The average length of sequences in 600 families ranges between 27 and 811, while the median is 312.

To predict structurally disordered regions in all protein sequences we used VSL2B predictor [11] since this was the most accurate disorder predictor at two consecutive protein structure prediction assessment community-wide experiments (CASP 6-7). We found that 18% of residues in the constructed dataset were predicted to belong to SDRs.

2.2 An iterative procedure for estimation of a 40x40 substitution matrix

Modifications of Needleman-Wunsch and Smith-Waterman algorithms (global and local pairwise sequence alignment) for use with extended alphabet and an expanded 40x40 substitution matrix were fairly straightforward. We implemented a multiple-sequence alignment algorithm based on ClustalW (as described in [7]) with necessary modifications. To save computation time we

pre-computed the all-to-all pairwise sequence identities using the Smith-Waterman algorithm and BLOSUM62 matrix (ClustalW uses a heuristic to estimate pairwise identities) and used the same guiding tree and weights for multiple-sequence alignment in all iterations.

We used the following iterative procedure for sequence alignment and estimation of the 40x40 substitution matrix:

1. Initialize the 40x40 matrix (as explained below).
2. Obtain multiple-sequence alignment for each family of sequences using the current matrix.
3. Calculate a new matrix from the alignments obtained in step 2.
4. Go back to step 2, unless the changes between iterations are negligible.

The first step of the iterative procedure initializes the matrix to a 40x40 matrix made up of four copies of BLOSUM62 substitution matrix (Figure 1). This means that in the first iteration of alignment, the disorder prediction information is ignored.

After the alignments are obtained in step 2, the new substitution matrix is calculated using the following procedure:

1. Initialize array for matrix M to zeros.
2. For each family of sequences:
   For each pair of sequences seq_i, seq_j, with weights w_i, w_j, for which $i < j$,
   For each pair of matched amino-acids from seq_i and seq_j, (excluding "matches" to gaps):
   increase the cell in the array corresponding to the two matched amino-acid by $w_1 w_2$.
3. Calculate matrix of amino acid pair frequencies $P = \left| p_{ij} \right|$ as $P = \left(M + M' \right) / 2 \sum_{i,j} m_{ij}$
4. Calculate frequencies for amino acids $q_i = \sum_j p_{ij}$
5. Calculate all scores using the formula: $score\left(a_i, a_j\right) = 2 \log_2 \left(p_{ij} / q_i q_j \right)$

The value of constant $C = 2$ is the same as in the calculation of the BLOSUM62 matrix, so the same gap penalty values can be used.

2.3 Experiments

All experiments with the described iterative procedure were performed with the default values for gap penalties in BLAST algorithm: 11 for gap opening and 1 for gap extension. In the main experiment we used the whole dataset to obtain a 40x40 substitution matrix.

To test the stability of our iterative procedure with respect to the choice of the dataset, we also run it six times with six different subsets of the dataset, each time randomly selecting only half of the sequence families. If the procedure is stable we expect to obtain six similar matrices.

As a control experiment, we modified the dataset by assigning randomly generated numbers instead of disorder predictions (we draw random numbers from a similar distribution as the values of disorder predictions). By comparing the matrices obtained in the main and control experiment, we were able to identify which properties of the matrix obtained in the main experiment are specific to structural disorder and were not obtained by pure chance.

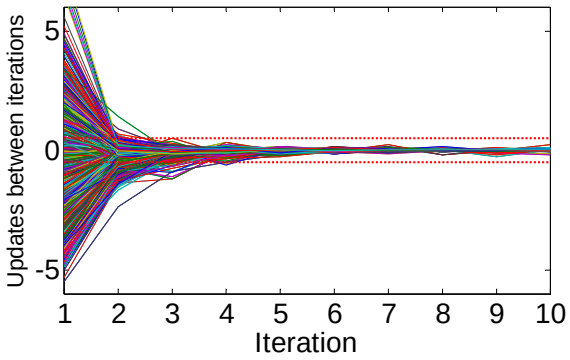

Figure 2. Convergence of the substitution matrix estimation procedure: relative updates for all 400 matrix elements are shown for the first 10 iterations. Horizontal lines are at $y = \pm 0.5$.

3. RESULTS

Our convergence criterion for the iterative procedure used to estimate the 40x40 substitution matrix is that the absolute values of updates for all parameters in the matrix fall below 0.5. This relaxed criterion is due to the fact that in applications the values in the matrix are usually rounded to the nearest integers to allow usage of integer arithmetic. The substitution matrix estimation procedure converged in five iterations as illustrated in Figure 2.

The 40x40 matrix obtained in the main experiment with the whole dataset is displayed in Table 1 (at the end of the Reference section). We compare the values in the obtained matrix with the values in the initial BLOSUM62 matrix in Figure 3.

We checked the stability of the iterative procedure by examining the distribution of std. deviations of six values obtained for each matrix element in the experiments repeated with six random subsets of the dataset (each subset contains 300 randomly selected sequence families, i.e. half of the dataset). Substitution matrices obtained in these six experiments were fairly similar with standard deviation for 85% of matrix elements smaller than 0.5 (histogram omitted for lack of space). The greatest instability is observed for scores related to least frequent amino acid types in disordered

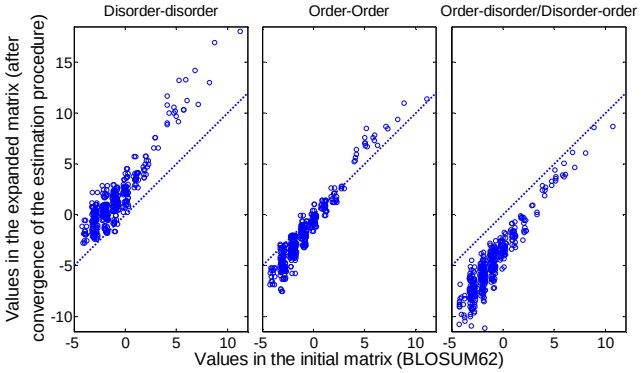

Figure 3. Comparison of values in the 20x20 submatrices of the 40x40 substitution matrix (obtained with our iterative procedure) versus the corresponding values in the initial matrix (BLOSUM62).

regions. This is expected, since $\log_2(p_{ij} / q_i q_j)$ is least stable for small values of p_{ij}, q_i and q_j.

For the 40x40 matrix obtained in the control experiment with the randomized dataset, we compare its four 20x20 submatrices. These four submatrices were practically identical, with 0.305 as the highest standard deviation for four related elements in these submatrices, and with standard deviation smaller than 0.1 for 85.5% of 400 submatrix positions. We also compared the four submatrices of the matrix obtained in the control experiment with the order-order submatrix of the substitution matrix obtained in the main experiment. The differences follow a distribution similar to a normal distribution with $\mu = .24$ and $\sigma = .16$, meaning that although the submatrices are very close, the scores are slightly higher in the order-order submatrix of the expanded substitution matrix from the main experiment.

4. DISCUSSION

The iterative procedure for estimation of the 40x40 substitution matrix that we described in this paper is an effective way of overcoming the lack of ground-truth alignments. The resulting substitution matrix is the fixed point of the mapping defined by steps 2 and 3 of the procedure. It also has the property that it both produces the alignments in step 2, and it is derived from the same alignments.

In the obtained expanded substitution matrix we observed substantial differences between the scores assigned to alignment of disordered-disordered, ordered-ordered and ordered-disordered pairs of amino acids. These differences provide further evidence that evolutionary rates in disordered and ordered regions of proteins are different and that BLOSUM62 and other matrices are not appropriate for alignment of SDPs. In contrast to BLOSUM62 matrix that tends to penalize matching of non-identical amino acids, our expanded matrix tends to assign higher scores (or at least smaller penalties) to the matching of non-identical amino acids in the disordered regions, where due to higher evolutionary rate such mismatches are more likely to occur in nature. The scores for alignment of ordered regions of two sequences in our expanded matrix are similar to scores assigned by the BLOSUM62 matrix. Finally, our matrix assigns the lowest scores (or more precisely: highest penalties) for matching amino acids in ordered regions in one sequence to amino acids in disordered regions in another sequence. This is consistent with the conservation of position and extent of disordered regions in homologous sequences.

The experiments with the random subsets of the dataset showed that the procedure is stable with respect to the selection choice of the protein sequences in the dataset (as long as the selection is done randomly). The results also emphasize the importance of using a large dataset. Furthermore, the results of the experiment with the randomized dataset showed that the differences between four 20x20 submatrices observed in the main experiment were not obtained by chance and that they clearly come from the differences between evolutionary rate in ordered and disordered regions of proteins.

We are currently running extensive testing of the iterative procedure with various values of gap opening (from 5 to 15) and extension penalties (0.5, 1, 2). In the matrices that we obtained so far for several combinations of gap penalties we found similar differences between 20x20 submatrices as was the case for the original experiment with 11/1 gap opening/extension penalties.

The 40x40 substitution matrix is ready to be used with the modified versions of local and global pairwise alignment algorithms, as well as with the modified version of multiple-sequence alignment algorithm. The only preprocessing required for this algorithm is the application of disorder predictor on sequences to be aligned.

The ultimate test for our proposed approach to protein sequence alignment will be its comparison with currently available alignment tools in real applications. Since aligning a query sequence against some large database of sequences is only feasible with heuristic-based algorithms such as BLAST, we are in process of implementing a modification of BLAST, which by itself is a very involved algorithm. Another option that we are currently exploring is development of a scheme that involves the original BLAST with appropriate pre- and post-processing. For this we are relying on PSI-BLAST, as it allows using a Position-Specific Scoring Matrix (PSSM) as an input.

5. ACKNOWLEDGMENTS

This work was supported by the grant R56 LM007688-05A1 from the National Institute of Health.

6. REFERENCES

[1] Henikoff, S. 1992. Amino Acid Substitution Matrices from Protein Blocks. PNAS 89: 10915–10919.

[2] Dunker, A. K. , Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. 2001. Intrinsically disordered protein. J Mol Graph Model 19, 26-59.

[3] Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. 2001. Sequence complexity of disordered protein. Proteins 42, 38-48.

[4] Brown, C.J., Takayama, S., Campen, A., Vise, P., Marshall, T., Oldfield, C.J., and Dunker, A.K. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. 55: 102-107.

[5] Radivojac P., Obradovic Z., Brown C.J., and Dunker A.K. 2002. Improving sequence alignments for intrinsically disordered proteins. Pac Symp Biocomput. 589-600.

[6] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J Mol Biol 215 (3): 403–410.

[7] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31 (13): 3497-3500.

[8] Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. 1978. A model of Evolutionary Change in Proteins, Atlas of protein sequence and structure (volume 5, supplement 3 ed.), Nat. Biomed. Res. Found., p. 345-358.

[9] Henikoff, J.G., and Henikoff, S. 1996. Blocks database and its applications. Methods Enzymol. 1996;266:88-105.

[10] Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. 2002. Intrinsic disorder and protein function. Biochemistry 41, 6573-6582.

[11] Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. 2006. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7, 208.

[12] Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48 (3): 443-53.

[13] Smith, T.F., Waterman, M.S. 1981. Identification of Common Molecular Subsequences. J Mol Biol 147: 195–197.

Table 1. The 40x40 substitution matrix obtained with our iterative procedure (initial matrix: BLOSUM62; gap opening penalty: 11; gap extension penalty 1). The matrix is divided into four 20x20 submatrices, as explained in the Introduction and Figure 1. The values are rounded to the nearest integer.

		Order																				Disorder																			
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	C	11	0	-1	-5	0	-3	-3	-5	-5	-4	-3	-4	-5	-2	-2	-2	-1	-2	-2	-3	9	-3	-4	-7	-3	-6	-6	-7	-7	-6	-5	-5	-7	-7	-6	-5	-4	-5	-5	-5
	S	0	6	1	-1	1	-1	0	-1	-1	-1	-2	-1	-1	-3	-4	-4	-3	-4	-3	-5	-6	2	-2	-4	-3	-4	-3	-3	-4	-4	-5	-4	-4	-7	-8	-8	-7	-8	-8	-9
	T	-1	1	7	-2	0	-3	-1	-2	-2	-1	-2	-2	-1	-1	-2	-3	-1	-4	-4	-4	-6	-1	3	-5	-3	-6	-3	-4	-4	-4	-4	-4	-4	-5	-5	-6	-4	-7	-7	-9
	P	-5	-1	-2	8	-1	-4	-3	-2	-2	-2	-3	-3	-2	-5	-5	-5	-4	-6	-5	-5	-9	-3	-5	5	-4	-6	-6	-4	-4	-5	-5	-6	-5	-8	-8	-8	-7	-9	-9	-10
	A	0	1	0	-1	5	-1	-2	-2	-1	-1	-3	-2	-2	-2	-2	-3	-1	-3	-4	-4	-5	-2	-3	-3	2	-3	-4	-4	-3	-4	-5	-4	-4	-5	-6	-5	-4	-7	-7	-8
	G	-3	-1	-3	-4	-1	7	-1	-2	-3	-3	-4	-3	-3	-5	-7	-6	-5	-6	-6	-6	-9	-4	-7	-7	-4	4	-4	-4	-5	-6	-6	-5	-6	-9	-11	-10	-9	-10	-9	-11
	N	-3	0	-1	-3	-2	-1	8	1	-1	0	1	-1	0	-4	-6	-5	-5	-5	-3	-5	-8	-2	-3	-6	-4	-3	4	-1	-3	-3	-2	-4	-3	-7	-8	-8	-8	-8	-6	-9
O	D	-5	-1	-2	-2	-2	-2	1	7	2	0	-1	-2	-1	-6	-8	-7	-6	-7	-5	-7	-11	-3	-4	-5	-4	-4	-1	4	0	-3	-4	-5	-4	-9	-10	-9	-8	-9	-8	-10
r	E	-5	-1	-2	-2	-1	-3	-1	2	7	2	-1	-1	1	-4	-6	-5	-4	-6	-4	-5	-10	-3	-4	-4	-3	-5	-3	0	3	-2	-4	-3	-2	-7	-8	-7	-6	-9	-8	-10
d	Q	-4	-1	-1	-2	-1	-3	0	0	2	8	1	1	1	-2	-4	-3	-4	-5	-3	-4	-8	-3	-4	-4	-3	-5	-3	-2	-1	4	-2	-2	-1	-5	-7	-5	-6	-7	-6	-8
e	H	-3	-2	-2	-3	-3	-4	1	-1	-1	1	9	0	-1	-3	-5	-4	-4	-2	1	-2	-7	-4	-5	-6	-5	-6	-3	-3	-4	-2	6	-3	-4	-7	-8	-6	-8	-5	-3	-6
r	R	-4	-1	-2	-3	-2	-3	-1	-2	-1	1	0	7	3	-3	-5	-4	-4	-5	-3	-3	-8	-4	-4	-5	-4	-5	-4	-4	-3	-2	-3	4	0	-6	-7	-6	-6	-8	-6	-7
	K	-5	-1	-1	-2	-2	-3	0	-1	1	1	-1	3	7	-3	-5	-4	-4	-6	-4	-5	-8	-3	-4	-5	-4	-5	-3	-3	-2	-2	-4	0	3	-7	-7	-6	-6	-8	-7	-10
	M	-2	-3	-1	-5	-2	-5	-4	-6	-4	-2	-3	-3	-3	8	1	2	0	0	-2	-2	-7	-5	-4	-8	-5	-8	-6	-7	-6	-4	-6	-5	-5	4	-3	-1	-3	-4	-5	-6
	I	-2	-4	-2	-5	-2	-7	-6	-8	-6	-4	-5	-5	-5	1	6	1	3	-1	-3	-3	-6	-6	-4	-7	-5	-9	-7	-8	-7	-6	-6	-6	-6	-2	3	-1	0	-3	-5	-7
	L	-2	-4	-3	-5	-3	-6	-5	-7	-5	-3	-4	-4	-4	2	1	5	0	0	-2	-2	-6	-6	-5	-7	-5	-9	-8	-8	-7	-5	-5	-5	-6	-2	-2	2	-2	-2	-4	-5
	V	-1	-3	-1	-4	-1	-5	-5	-6	-4	-4	-4	-4	-4	0	3	0	6	-1	-3	-4	-5	-5	-3	-6	-3	-7	-6	-7	-5	-6	-6	-6	-6	-3	0	-2	3	-4	-5	-7
	F	-2	-4	-4	-6	-3	-6	-5	-7	-6	-5	-2	-5	-6	0	-1	0	-1	8	3	1	-7	-6	-7	-8	-6	-9	-8	-8	-8	-8	-3	-7	-8	-4	-4	-2	-4	5	1	-2
	Y	-2	-3	-4	-5	-4	-6	-3	-5	-4	-3	1	-3	-4	-2	-3	-2	-3	3	9	1	-6	-6	-6	-8	-6	-8	-5	-6	-6	-5	-1	-5	-6	-5	-5	-4	-6	1	6	-2
	W	-3	-5	-4	-5	-4	-6	-5	-7	-5	-4	-2	-3	-5	-2	-3	-2	-4	1	1	11	-6	-7	-7	-8	-7	-8	-8	-9	-8	-6	-4	-5	-7	-5	-6	-4	-6	-1	0	9
	C	9	-6	-6	-9	-5	-9	-8	-11	-10	-8	-7	-8	-8	-7	-6	-6	-5	-7	-6	-6	17	2	1	-1	1	1	1	-2	-3	0	0	1	-1	-2	0	0	1	1	2	3
	S	-3	2	-1	-3	-2	-4	-2	-3	-3	-3	-4	-4	-3	-5	-6	-6	-5	-6	-6	-7	2	9	4	2	3	2	4	2	1	2	1	1	1	-1	0	0	1	0	0	-1
	T	-4	-2	3	-5	-3	-7	-3	-4	-4	-4	-5	-4	-4	-4	-4	-5	-3	-7	-6	-7	1	4	10	2	4	1	3	2	1	2	1	1	2	1	2	1	3	0	0	-2
	P	-7	-4	-5	5	-3	-7	-6	-5	-4	-4	-6	-5	-5	-8	-7	-7	-6	-8	-8	-8	-1	2	2	11	3	0	1	1	0	2	1	0	0	-2	0	1	1	-1	-2	-1
	A	-3	-3	-3	-4	2	-4	-4	-4	-3	-3	-5	-4	-4	-5	-5	-5	-3	-6	-6	-7	1	3	4	3	9	2	2	2	2	3	1	1	2	0	1	1	3	0	0	-1
	G	-6	-4	-6	-6	-3	4	-3	-4	-5	-5	-6	-5	-5	-8	-9	-9	-7	-9	-8	-8	1	2	1	0	2	10	2	2	1	0	0	2	0	-2	-2	-2	0	-2	-2	0
D	N	-6	-3	-3	-6	-4	-4	4	-1	-3	-3	-3	-4	-3	-6	-7	-8	-6	-8	-5	-8	1	4	3	1	2	2	11	4	2	3	3	1	3	-1	0	-1	0	-1	1	-2
i	D	-7	-3	-4	-4	-4	-4	-1	4	0	-2	-3	-4	-3	-7	-8	-8	-7	-8	-6	-9	-2	2	2	1	2	2	4	10	5	2	2	0	1	-2	-1	-2	0	-2	-1	-3
s	E	-7	-4	-4	-4	-3	-5	-3	0	3	-1	-4	-3	-2	-6	-7	-7	-5	-8	-6	-8	-3	1	1	0	2	1	2	5	9	4	1	1	3	-1	-1	-1	0	-2	-2	-2
o	Q	-6	-4	-4	-5	-4	-6	-3	-3	-2	4	-2	-2	-2	-4	-6	-5	-6	-8	-5	-6	0	2	2	2	3	0	3	2	4	11	4	3	3	0	0	1	1	-1	0	1
r	H	-5	-5	-4	-5	-5	-6	-2	-4	-4	-2	6	-3	-4	-6	-6	-5	-6	-3	-1	-4	0	1	1	1	1	0	3	2	1	4	13	3	1	-1	0	0	0	2	4	1
d	R	-5	-4	-4	-6	-4	-5	-4	-5	-3	-2	-3	4	0	-5	-6	-5	-6	-7	-5	-5	1	1	1	0	1	2	1	0	1	3	3	10	5	-1	0	0	0	-2	-1	2
e	K	-7	-4	-4	-5	-4	-6	-3	-4	-2	-1	-4	0	3	-5	-6	-6	-6	-8	-6	-7	-1	1	2	0	2	0	3	1	3	3	1	5	10	-1	0	-1	0	-2	-1	-2
r	M	-7	-7	-5	-8	-5	-9	-7	-9	-7	-5	-7	-6	-7	4	-2	-2	-3	-4	-5	-5	-2	-1	1	-2	0	-2	-1	-2	-1	0	-1	-1	-1	13	4	4	3	2	0	0
	I	-6	-8	-5	-8	-6	-11	-8	-10	-8	-7	-8	-7	-7	-3	3	-2	0	-4	-5	-6	0	0	2	0	1	-2	0	-1	-1	0	0	0	0	4	12	5	7	4	2	1
	L	-5	-8	-6	-8	-5	-10	-8	-9	-7	-5	-6	-6	-6	-1	-1	2	-2	-2	-4	-4	0	0	1	1	1	-2	-1	-2	-1	1	0	0	-1	4	5	10	4	5	2	2
	V	-4	-7	-4	-7	-4	-9	-8	-8	-6	-6	-8	-6	-6	-3	0	-2	3	-4	-6	-6	1	1	3	1	3	0	0	0	0	1	0	0	0	3	7	4	11	3	1	0
	F	-5	-8	-7	-9	-7	-10	-8	-9	-9	-7	-5	-8	-8	-4	-3	-2	-4	5	1	-1	1	0	0	-1	0	-2	-1	-2	-2	-1	2	-2	-2	2	4	5	3	13	8	6
	Y	-5	-8	-7	-9	-7	-9	-6	-8	-8	-6	-3	-6	-7	-5	-5	-4	-5	1	6	0	2	0	0	-2	0	-2	1	-1	-2	0	4	-1	-1	0	2	2	1	8	14	6
	W	-5	-9	-9	-10	-8	-11	-9	-10	-10	-8	-6	-7	-10	-6	-7	-5	-7	-2	-2	9	3	-1	-2	-1	-1	0	-2	-3	-2	1	1	2	-2	0	1	2	0	6	6	18