Accuracy of Protein Flexibility Predictions

Mauno Vihinen,[1] Esa Torkkila,[2] and Pentti Riikonen[2]
[1]Department of Biochemistry, University of Turku, SF-20500 Turku and Turku Centre for Biotechnology, University of Turku, SF-20520 Turku and [2]Department of Computer Science, University of Turku, SF-20520 Turku, Finland

ABSTRACT **Protein structural flexibility is important for catalysis, binding, and allostery. Flexibility has been predicted from amino acid sequence with a sliding window averaging technique and applied primarily to epitope search. New prediction parameters were derived from 92 refined protein structures in an unbiased selection of the Protein Data Bank by developing further the method of Karplus and Schulz (Naturwissenschaften 72:212–213, 1985). The accuracy of four flexibility prediction techniques was studied by comparing atomic temperature factors of known three-dimensional protein structures to predictions by using correlation coefficients. The size of the prediction window was optimized for each method. Predictions made with our new parameters, using an optimized window size of 9 residues in the prediction window, were giving the best results. The difference from another previously used technique was small, whereas two other methods were much poorer. Applicability of the predictions was also tested by searching for known epitopes from amino acid sequences. The best techniques predicted correctly 20 of 31 continuous epitopes in seven proteins. Flexibility parameters have previously been used for calculating protein average flexibility indices which are inversely correlated to protein stability. Indices with the new parameters showed better correlation to protein stability than those used previously; furthermore they had relationship even when the old parameters failed.**

Key words: dynamics, flexibility index, protein stability, antigenic regions, epitopes

INTRODUCTION

Protein molecules are dynamic being in constant motion. Structural flexibility is essential for activity but, on the other hand, structural stability requires rigidity.[1–3] Flexible regions are found in catalytic sites,[4–7] binding sites,[8] antigenic regions,[9] sites susceptible for proteolytic cleavage,[10] allosteric hinge sites,[11] etc. Proteins with similar functions have similar excess of flexibility in their optimum reaction conditions.[2,4]

The core of a globular protein is relatively tightly packed. Surface residues are generally more mobile due to fewer stabilizing interactions. Exposed surface loops are the most flexible and show the largest sequence variation. The time scale of protein mobility is very wide, the fastest vibrations and motions requiring only 10^{-14} to 10^{-13} s. Mobility can be simulated with molecular dynamics. Although the simulations are relatively short plenty of valuable information is available. The flexible regions can be predicted using less accurate methods even without structural information.

Three techniques have been used for predicting protein flexibility from amino acid sequence. The methods of Karplus and Schulz[12] (KS) and of Bhaskaran and Ponnuswamy[13] (BP) are based on parameters derived from three-dimensional structures. Ragone et al.[14] (R) base their approach on a combination of hydropathy predictions and amino acid volumes. Flexibility analysis can be used to search for the most mobile and thus possibly also the surface residues in a sequence, which are thought to represent epitopes. For vaccine production, it would be of great value to be able to predict the antigenic regions of a protein from its sequence. Flexibility predictions have been used in searching for continuous epitopes from amino acid sequences.[12,14,15] Other epitope prediction methods include hydropathy,[16,17] β-turn propensity,[18] and joint prediction of hydropathy, surface accessibility, flexibility, and secondary structure.[19] Stern[20] has recently reviewed the methods.

The KS method uses normalized *B*-values of C_α-atoms in 31 protein structures. Here we have extended the flexibility prediction by analyzing all the backbone atoms of 92 well refined structures from an unbiased selection of PDB.[21] To test the applicability of the predictions they were compared to ex-

Abbreviations: BP, flexibility according to Bhaskaran and Ponnuswamy; KS, according to Karplus and Schulz; R, according to Ragone et al.; VTR, according to parameters derived here; PDB, Protein Data Bank.

Received October 5, 1993; revision accepted February 1, 1994.

Address reprint requests to Mauno Vihinen at his present address: Center for Structural Biochemistry, Karolinska Institute, Novum, S-14157 Huddinge, Sweden.

perimental B-values. Their use in predicting antigenic regions was studied with proteins for which locations of continuous epitopes have been determined. Increased hydrophobicity and decreased flexibility have been shown to be the main stabilizing principles in thermostable proteins.[22] Previously we have shown inverse correlation between thermal stability and structural flexibility by calculating flexibility indices from normalized B-values for amino acid sequences.[2] Even a better and more accurate correlation to stability was noticed when flexibility indices were calculated with the new parameters.

METHODS

Entries for high resolution structures containing B-values were taken from the unbiased selection of Protein Data Bank[21] (PDB) because there are a lot of redundant data in PDB. Only 78 of the original 102 entries could be used in this analysis because of missing or incomplete B-values or sequence information. Some of the chosen entries contained several proteins, so finally there were 92 different structures. The PDB entries were 1bp2, 1ccr, 1cla, 1cse (2 chains), 1ctf, 1eca, 1fc2 (2 chains), 1gcr, 1gd1, 1gox, 1gp1, 1hoe, 1i1b, 1ldm, 1lz1, 1mbd, 1nxb, 1pcy, 1phh, 1prc (4 chains), 1r69, 1sgt, 1sn3, 1tnf, 1ubq, 1utg, 1wsy (2 chains), 2aza, 2cab, 2ccy, 2cdv, 2ci2, 2cpp, 2cts, 2cyp, 2fb4 (2 chains), 2gbp, 2gn5, 2hhb (2 chains), 2hla (2 chains), 2hmg (2 chains), 2lbp, 2lh2, 2ltn (2 chains), 2mhr, 2ovo, 2paz, 2pfk, 2rnt, 2rsp, 2sga, 2sod, 2ts1, 2wrp, 3adk, 3gap, 3grs, 3ins (2 chains), 3lzm, 3rn3, 3tln, 451c, 4cha, 4fd1, 4hvp, 4pep, 4xia, 5at1 (2 chains), 5cpa, 5pti, 5rxn, 6acn, 7api (2 chains), 8adh, 8cat, 8dfr, 9pap, 9wga. Normalized B-values derived from the unbiased structures were used both for flexibility prediction and the calculation of flexibility indices. Computer programs were developed to be compatible with the GCG program suite.[23]

Calculation of Normalized *B*-Values

The selection of 92 unbiased protein structures was used to derive normalized B-values. Temperature factors of the backbone atoms N, C_α, C, and O were taken from the PDB.[24] The Karplus and Schulz[12] approach of determining normalized B-values was repeated with our extended database. The threshold values are those previously used. The B-values of each protein were normalized so that the mean was 1.0 and the root mean square deviation 0.3. Based on its deviation from the mean, each residue type was defined as flexible or rigid. Those with average B_{norm} values below 1.0 were denoted as rigid. In the next step normalized B-values were determined for each residue type when surrounded by none, one or two rigid neighbours to obtain B_{norm0}, B_{norm1}, and B_{norm2} tables, respectively. Because chain termini are usually very flexible and could have caused bias, three N- and C-terminal residues were omitted from each structure.

Programs for Flexibility Prediction

Program FLEX was implemented for flexibility predictions with our new B_{norm} tables and with the parameters of Bhaskaran and Ponnuswamy[13] and Ragone et al.[14] The antigenic index of Jameson and Wolf[19] was also included. The sequence can be read either from a PDB or a GCG file. The predictions are based on a sliding window averaging technique. The optimized window size for each technique is used: five for R, seven for BP, and nine residues for our parameters and those of KS. The propensities for the residues inside the window are summed up and given for the residue in the middle of the window. The weighting of residues inside the window is 0.25, 0.4375, 0.625, 0.8125, 1, 0.8125, 0.625, 0.4375, and 0.25 from left to right in techniques using B_{norm} values but has a constant value of 1 in the methods of BP and R. The flexibilities of KS and ours are calculated as follows. First the number of rigid neighbors around each residue is determined. Then the neighbor correlated weighted propensities from B_{norm} tables are summed and given for the middlemost residue after which the window is shifted by one residue. The results can be presented with the program FLEXPLOT on several graphics devices. Experimental B-values are shown for the backbone atoms of proteins in PDB entries.

Testing Accuracy of the Flexibility Predictions

The accuracy of the different flexibility prediction methods was studied by determining correlation coefficients. The B-values for each of the proteins were compared to predicted flexibilities by calculating correlation coefficients. Many PDB structures contain one or just few highly flexible residues due to, e.g., lattice disorders. To see if the high peaks might bias the analysis, the B-values of residues in each protein were scaled from 0 to 100%. If only one residue had flexibility higher than 80 or 90%, its value was reduced to that of the second highest residue and the analysis was repeated until there were residues also on intervals 80 to 90% and/or 90 to 100%. The correlation coefficients were determined for the entries both when smoothed by sieving the high peaks and when untreated.

Optimization of the Flexibility Prediction Techniques

The flexibility prediction techniques use the sliding window averaging technique. The only adjustable parameter in the R and BP methods is the width of the window, i.e., number of consecutive residues used in the prediction at a time. In the method of KS the window was originally fixed to seven residues but the residues inside the window had different weighting depending on their location within the window. The length of the window was optimized for all four prediction techniques by determining correlation coefficients and maximizing information contents with window lengths 5 to 15 residues. In addition also the effect of residue weighting was studied by giving the weight of 0.25 for the first and last residues in the window and 1.0 for the middlemost. The weights of the others were at equal spacing between these two values.

Calculation of Flexibility Indices

Atomic temperature factors (B-values) obtained during crystal structure determination are a measure of the flexibility of the residues in the protein. We have used normalized B-values to calculate average flexibility indices for the whole protein molecule. Since the flexibility of a residue is dependent on the nature of neighboring residues, three parameter tables are used. There have been two ways to calculate average flexibility indices.[2] The F index is calculated from

$$F = \sum_{i=2}^{n-1} B_{\mathrm{nc},i}/(n-2)$$

where n is the number of residue and B_{nc} is neighbor correlated normalized B-value for the residue type. Another equation, F_7, gives different emphasis for the chain termini

$$F_7 = \sum_{i=8}^{n-7} f_i/(n-8)$$

where $f_i = [B_{\mathrm{nc},i} + 0.75(B_{\mathrm{nc},i-1} + B_{\mathrm{nc},i+1}) + 0.5(B_{\mathrm{nc},i-2} + B_{\mathrm{nc},i+2}) + 0.25(B_{\mathrm{nc},i-3} + B_{\mathrm{nc},i+3})]/4$. Now that window size nine was found to be optimal in predictions with normalized B-values a new equation was determined

$$F_9 = \sum_{i=10}^{n-9} f_i/(n-10)$$

where $f_i = [B_{\mathrm{nc},i} + 0.8125(B_{\mathrm{nc},i-1} + B_{\mathrm{nc},i+1}) + 0.625(B_{\mathrm{nc},i-2} + B_{\mathrm{nc},i+2}) + 0.4375(B_{\mathrm{nc},i-3} + B_{\mathrm{nc},i+3}) + 0.25(B_{\mathrm{nc},i-4} + B_{\mathrm{nc},i+4})]/5.25$.

RESULTS AND DISCUSSION

New Flexibility Parameters

Three methods have been used to predict protein structural flexibility from sequences.[12–14] The parameters for the KS- and BP-techniques were derived from known 3D structures, whereas those for the R-technique are combined from other predictions. A limited set of 31 structures was used in the KS method to determine prediction parameters, whereas BP had only 19 proteins. We have extended the analysis to 92 refined structures. We reimplemented the KS algorithm because we found it gave the most accurate predictions. All these techniques use a sliding window averaging technique; parameters are summed for a stretch of amino acids within a window which is shifted by one residue at a time. In the KS method residues have coefficients dependent on the location within the window, thus the contribution of a residue to the prediction value depends on its distance from the middle of the window. Claverie and Daulmerie[25] argue that smoothing of the prediction curves by weighting is advantageous, since pattern recognition is easier and the irregular variation of values is damped. The smoothing is also better for detecting local maxima which are of importance, e.g., in epitope analysis. von Heijne[26] has used a related trapezoid weighting scheme in analysis of membrane spanning segments.

The proteins for calculating normalized B-values were taken from an unbiased selection of the PDB.[21] Normalized B-values, hereafter VTR parameters, were calculated from the 92 structures (Table I). The major difference to those of KS is that we have 11 rigid residues instead of 10. Threonine is classified as a rigid amino acid, because its average B_{norm} is below 1. The order and values of residues have changed. Glycine is generally considered to be the most flexible amino acid. It has the highest value both in BP and R tables but not in the KS table. In our analysis it is found to be flexible but there are still seven more flexible residue types. This might be because the more flexible residues, which are all charged or polar except for proline, appear mainly on surface whereas glycine is also found in the protein interior. As the normalized B-values are averages the restricted mobilities of buried glycine residues may reduce the overall value. Another explanation might be frequent occurrence in tight turns having restricted mobility. The values for glycine are the most neighbour dependent. When surrounded by one or two rigid residues it is among the most flexible residues.

The new B_{norm} values were used in flexibility prediction. If the sequence in program FLEX is read from PDB file B-values are averaged for the backbone atoms. The predictions and the B-values are presented with program PLOTFLEX. For the plots the values of BP and R tables were normalized to be from 0 to 1. In the original R parameterization the most flexible residue had the lowest value thus the numbers were inversed to be comparable to the others.

Accuracy of the Flexibility Prediction Techniques

The accuracy of the techniques was tested with correlation coefficients method. The prediction window was adjusted from 5 to 15 residues and the prediction accuracy was followed when the highest B-value peaks were either smoothed or not. The means of the correlation coefficients over the 92 proteins in Table II shows that the optimal window in

TABLE I. Neighbor Correlated Normalized Flexibility Parameters of the 92 Protein Structures

			B_{norm0}		B_{norm1}		B_{norm2}	
Resid.	Count	$B_{\text{norm,avr}}$	Count	Value	Count	Value	Count	Value
W	264	0.904	51	1.186	60	0.938	153	0.796
C	333	0.906	68	1.196	79	0.939	186	0.785
F	708	0.915	159	1.247	154	0.934	395	0.774
I	926	0.927	208	1.241	213	0.977	505	0.776
Y	646	0.929	144	1.199	165	0.981	337	0.788
V	1297	0.931	296	1.235	325	0.968	676	0.781
L	1505	0.935	346	1.234	365	0.982	794	0.783
H	457	0.950	112	1.279	121	0.967	224	0.777
M	349	0.952	85	1.269	75	0.963	189	0.806
A	1499	0.984	432	1.315	338	0.994	729	0.783
T	1057	0.997	300	1.324	270	0.998	487	0.795
R	764	1.008	225	1.310	186	1.026	353	0.807
G	1529	1.031	491	1.382	359	1.018	679	0.784
Q	674	1.037	213	1.342	161	1.041	300	0.817
S	1171	1.046	378	1.381	279	1.025	514	0.811
N	794	1.048	255	1.380	221	1.022	318	0.799
P	857	1.049	295	1.342	201	1.050	361	0.809
D	1011	1.068	371	1.372	226	1.022	414	0.822
E	1027	1.094	396	1.376	253	1.052	378	0.826
K	1038	1.102	420	1.367	278	1.029	340	0.834

BP method was seven residues and five in R. We can see that the R method is the overall poorest technique. This can be understood from the origin of the parameters which were obtained by multiplying residue hydrophobicities by volumes without using any structural analysis. The three methods giving better correlation coefficients are based on three-dimensional structures.

The optimization of the predictions with the VTR and KS parameters required weighting of the residues in the window. The window sizes tested were from 5 to 15 residues. The weights of the outermost residues were 0.25 and 1 for that in the middle. The value 0.25 was chosen to give some emphasis also for the ends of windows. The optimal window size was nine for both VTR and KS techniques (Table II), the latter of which previously used seven residues. The correlation coefficients with the optimized prediction techniques for all the 92 protein structures with all the four techniques are as follows: VTR 0.3304, KS 0.3356, BP 0.2428, and R 0.1659. Clearly the best results were obtained with the VTR and KS methods, the other two being much poorer. The predictive power varies greatly in each technique. The best correlation coefficients are close to 0.8 whereas the poorest values are close to 0. All the tested methods predicted poorly some proteins. The results show that in the KS method the previously used window of 7 residues is not optimal. The best results are obtained with nine consecutive amino acids. The new parameters were better than those of KS with short (5 to 7 residues) and large (15 residues) window but the differences are not significant.

Many proteins have one or only a few residues with very large B-values due to, e.g., static disorders in crystal lattice. These residues produce high peaks on B-value curves and might bias the parameters for those residues. This could be avoided by smoothing the curve, but no real effect on predictability was seen, e.g., in the case of the VTR method the unsmoothed value with window size 9 was 0.3302 while it was 0.3304 for the smoothed data. The same order of improvement was noticed also in the other three methods. Many of the highest peaks were already filtered away when the three N- and C-terminal residues of each protein were not included in the calculation of prediction parameters. This was done because the ends are known to be exceptionally flexible.

The use of backbone atoms was tested comparing predictability to parameters derived from C_α atoms of the 92 proteins. Correlation coefficients were determined for both parameter sets and the mean was found to be 0.330 for the backbone derived data with window 9 whereas is was 0.320 for the tables derived from C_α atoms. The improvement in the backbone-derived data is surprisingly small. The same result can be noticed when comparing the results of backbone parameters (VTR scale) to those of C_α parameters (KS table). It seems that in KS analysis there were enough data to bring the predictability with this sort of technique close to its maximum and our data for 17,906 amino acids did not change it much.

We tested the prediction methods also with structures not included in our data set; 38 randomly selected structures from a later version of PDB not having significant sequence similarity to the proteins used in the derivation of the parameters were

TABLE II. Optimization of Flexibility Prediction Windows*

Window	Prediction technique			
size	VTR	KS	BP	R
5	0.3158	0.3112	0.2345	0.1659
7	0.3266	0.3283	0.2428	0.1655
9	0.3304	0.3356	0.2387	0.1644
11	0.3280	0.3332	0.2219	0.1602
13	0.3204	0.3235	0.2092	0.1628
15	0.3125	0.3142	0.2030	0.1645

*The overall correlation coefficients for all the chosen 92 proteins were determined with different prediction window sizes.

analyzed. Here, too, predictions with the VTR and KS methods are giving best results (mean values 0.3359 and 0.3260, respectively), R and BP scales are clearly the worst ones (mean values 0.2460 and 0.2596, respectively). The accuracy of the predictions are of the same order as for the structures used to derive the parameters, but the differences are not significant. The VTR is somewhat better than the KS method. The most striking result is an increase in the predictability of the R method. VTR and KS parameters ar the best and the new parameters are somewhat more accurate.

The applicability of the flexibility predictions is shown for myoglobin in Figure 1. The VTR and KS plots resemble each other although the new scale is discriminating flexible and rigid regions more sharply, which is advantageous in searching for antigenic regions. The flexibilities of the two techniques follow quite well the shape of the B-value curves, although the predicted curves are smoother.

The flexibility predictions and experimental B-values could further be compared with the program MULTICOMP,[27] a multiple sequence comparing tool which can also be used for comparing predictions. Prior to this kind of analysis the B-values and flexibility propensities have to be normalized to express the same range of values. This approach has also been used to compare hydropathy predictions by comparing two different methods of predicting hydropathic character on the same protein.[28]

Prediction of Antigenic Sites

The protein surface serves as a template for numerous antibodies. Some of the epitopic regions are formed by consecutive residues. These regions have been determined for several proteins such as sperm whale myoglobin[16] (PDB entry 1 mbo), hen egg white lysozyme[16] (1lyz), tobacco mosaic virus protein[29] (2tmv), horse cytochrome c[16] (sequence entry ccho), bovine serum albumin[30] (a36401), rotavirus major outer-shell glycoprotein[31] (vs09_rotsl), and hepatitis B virus core protein[32] (nkvlah). The proteins contained although 31 continuous epitopes when the N- and C-terminal regions were omitted. Since one or the major applications of the flexibility predictions has been epitope search all four prediction techniques were used to locate antigenic regions in the seven proteins.

Each prediction technique was run with the optimized window sizes. Since there are no general rules to locate the antigenic regions from plots, areas having some sort of peak in the epitope region were considered to match. The VTR, KS, and R parameters predicted correctly 20 of the 31 epitopes which means 65% success ratio. The BP method was much poorer giving only 13 correct regions, 42% success. These figures might be reasonably good for this sort of simple method were there not also a high number of false positives. In Figure 1 we have included also the antigenic index,[19] which is specially made for epitope search. However, it was most often indicating some 60% of the sequence as highly antigenic, thus we did not consider that method at all.

Hydropathy profiles have generally been used for searching epitopes. The method of Hopp and Woods,[33] perhaps the most often used prediction technique for this purpose, was used to analyze the same proteins. There were 21 correctly predicted sites indicating no difference in accuracy to flexibility techniques. Because of the vague nature of the flexibility we could not calculate the ratios of correctly and wrongly predicted regions. Anyhow, it could be noted that by far the best methods for searching epitopes among the highest peaks in predictions are VTR and KS. They also predicted fewer false epitopes. The new parameters were better because they separated the peaks more clearly, which makes the interpretation of the results clearer and more accurate. The hydropathy predictions were made with program HYDRO.[34] Note that the hydrophilic regions are pointing down in the Hopp and Woods[33] prediction.

Flexibility Indices

The functional properties of a molecule are a compromise between flexibility and rigidity. The correlation between averaged flexibilities and protein thermal stability has been verified with flexibility indices calculated from the normalized B-values of KS.[2] Here we used VTR parameters to calculate also F indices. The values determined with KS parameters are shown for comparison. The differences in KS results to those previously published are due to a minor error in the routine for calculating F_7 in the previous work. Several groups of enzymes studied (Table III) indicated that the correlation to protein stability was even clearer with the new parameters. Indices calculated with VTR parameters show correlation also in alanine dehydrogenases, glucoamylases, serine proteases, and phosphoglycerate kinases, but not with those of KS. The flexibility indices are comparable for proteins having similar function and folding. Because they do not take into

TABLE III. Flexibility Indices of Some Proteins*

	Temperature			Our parameters		KS parameters		
Source	Optimum	Stability	T_m	F	F_9	F	F_9	Reference
Alanine dehydrogenase								
Bacillus sphaericus IFO3525		50%, 65°C, 5 min		1.0041	1.0035	0.9897	0.9897	35
Bacillus stearothermophilus IFO12550		50%, 85°C, 5 min		0.9958	0.9964	0.9903	0.9905	35
α-Amylase								
Bacillus subtilis 168		30%, 65°C, 10 min		1.0551	1.0548	1.0081	1.0080	
Streptomyces griseus IMRU 3570	42			1.0223	1.0217	1.0002	0.9998	36, 37
Bacillus amyloliquefaciens	50–60			1.0470	1.0484	1.0051	1.0054	
Apsergillus oryzae		70%, 50°C, 30 min		1.0125	1.0109	0.9978	0.9972	
B. stearothermophilus ATCC 12980	80			1.0178	1.0187	0.9957	0.9963	
B. stearothermophilus NZ-3		50%, 90°C, 2 h		1.0186	1.0196	0.9967	0.9973	
Bacillus licheniformis NCIB 8061		100%, 90°C, 2h		1.0295	1.0304	0.9971	0.9976	
β-Amylase								
Bacillus circulans NCIB 11033	50	77%, 57°C, 1 h		1.0400	1.0411	1.0044	1.0050	38
Clostridium thermosulfurogenes ATCC 33743		100%, 70°C, 1 h		1.0048	1.0055	0.9936	0.9939	39
β-Glucanase								
B. amyloliquefaciens		50%, 70°C, 4 min		1.0227	1.0190	0.9969	0.9954	40
Bacillus macerans		50%, 70°C, 9 min		1.0106	1.0124	0.9956	0.9950	40
Cyclodextrin glycosyltransferase								
Klebsiella pneumoniae M5a1		100%, 45°C, 15 min (+Ca)		1.0405	1.0395	1.0065	1.0058	
B. macerans IAM 1243	60	90%, 50°C, 15 min		1.0226	1.0221	1.0026	1.0026	
Bacillus circulans ATCC 21783	45	100%, 65°C, 30 min (+Ca)		1.0202	1.0198	0.9979	0.9978	41, 42
Ferredoxin								
Clostridium acidi-urici		22%, 70°C, 2h		0.9970	1.0048	0.9651	0.9702	
Clostridium tartarivorum		53%, 70°C, 2h		0.9622	0.9659	0.9659	0.9660	
Clostridium thermosaccharolyticum		90%, 70°C, 2 h		0.9625	0.9663	0.9677	0.9681	
Glucoamylase								
Schizosaccharomycopsis fibuligera HUT 7212	50			1.0475	1.0482	1.0068	1.0071	43
Schwanniomyces occidentalis ATCC 26076	52			1.0275	1.0279	0.9990	0.9991	44
Aspergillus awamori	70			1.0173	1.0184	1.0067	1.0068	45, 46
Inorganic pyrophosphatase								
Saccharomyces cerevisiae		20%, 50°C, 5 min		1.0405	1.0420	1.0044	1.0038	
Escherichia coli		35%, 90°C, 5 min		1.0273	1.0245	0.9921	0.9912	
Lactate dehydrogenase								
Lactobacillus casei DMS 20011		100%, 60°C, 5 min		1.0153	1.0143	0.9923	0.9917	
Bacillus psychrosaccharolyticus DSM 6	40/35			1.0144	1.0107	0.9904	0.9890	47, 48
Bacillus megaterium		100%, 48°C, 30 min		1.0130	1.0117	0.9906	0.9898	48, 49
B. subtilis X1	50–60	100%, 55°C, 30 min		1.0099	1.0108	0.9897	0.9898	
Bacillus caldotenax YT-G	60/70	100%, 65°C, 30 min		1.0110	1.0090	0.9870	0.9865	
Bacillus caldolyticus		100%, 70°C, 30 min		1.0037	1.0020	0.9805	0.9794	50, 51
B. stearothermophilus NCIB 8924	55/60–70	100%, 73°C, 30 min		1.0021	1.0003	0.9796	0.9786	
Thermus caldophilus GK 24		100%, 90°C, 60 min		0.9992	1.0015	0.9816	0.9821	52
Neutral protease								
B. amyloliquefaciens				1.0407	1.0449	1.0128	1.0134	53
B. subtilis				1.0384	1.0418	1.0143	1.0154	54
B. cereus DSM 3101		75%, 65°C, 20 min		1.0292	1.0295	1.0045	1.0042	55, 56
B. stearothermophilus CU21		80%, 65°C, 30 min		1.0232	1.0252	0.9957	0.9965	57, 58
B. caldolyticus YT-p	77	70%, 76°C, 30 min		1.0215	1.0234	0.9958	0.9961	59
Bacillus thermoproteolyticus		30%, 90°C, 30 min		1.0260	1.0273	0.9995	0.9991	58, 60
B. stearothermophilus MK232		45%, 90°C, 30 min		1.0260	1.0272	0.9998	0.9994	61, 62
Phosphoglycerate kinase								
S. cerevisiae			55	1.0338	1.0316	1.0004	0.9997	4, 63
Thermus thermophilus HB-8			>90	1.0314	1.0330	0.9898	0.9898	4, 64
Serine protease								
B. amyloliquefaciens	50–75			1.0340	1.0353	1.0041	1.0049	65, 66
Thermoactinomyces vulgaris	60–85	50%, 55°C, 40 min		1.0271	1.0271	1.0074	1.0076	67, 68
Thermus aquaticus YT-1	80	85%, 80°C, 3 h		1.0202	1.0197	1.0050	1.0041	69, 70

*If reference is not given it is mentioned in Vihinen.[2]

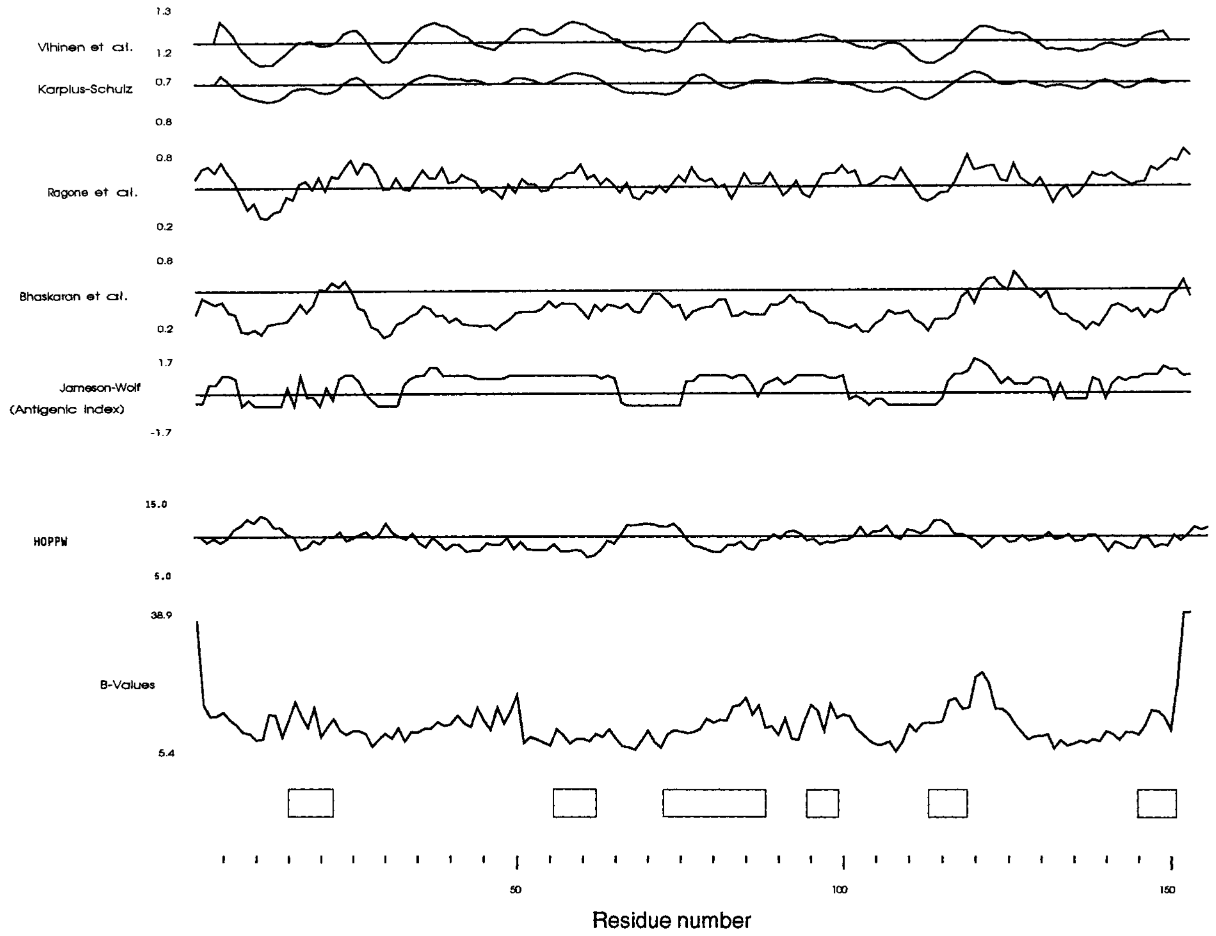

Fig. 1. Flexibility predictions, antigenic index, and experimental *B*-values of backbone atoms in sperm whale myoglogin (1mbo). The flexibility predictions were obtained with optimized prediction windows. Continuous epitopes are indicated with open boxes.

account all the stabilizing forces some discrepancy has been noticed.[2] This can still be seen in the case of the most stable α-amylases, ferredoxins, and neutral proteases. Ferredoxins with extra stabilizing ionic bonds have been discussed previously.[2] In neutral proteases the higher stability is presumably gained by the extra Ca^{2+} binding site. F_9 indices were determined also with KS parameters, but the results are not shown here because the difference to F_7 values were not higher than 0.0005, usually much less.

The flexibility indices calculated with the VTR parameters had more pronounced correlation to stability data than the KS parameters. Our values show correlation even when those of KS fail. This is presumably due to two reasons. Our structural database is larger. We also used the backbone information instead of C_α atoms. All the atoms in residues were not used because flexibility of side chains does not mean that the backbone is also flexible and the flexibility of the protein backbone is typical for surface regions and epitopes. Side chains can be rather mobile although the backbone is rigid. Because one of the major applications of the method will be to search for epitopes and mobile exposed regions only the backbone data were used. Another reason was that often the data for side chains are missing or are poorly determined.

CONCLUSIONS

New parameters were determined for prediction of protein flexibility. The applicability was studied by comparing atomic temperature factors of crystallographically determined proteins to predictions. The VTR and KS parameters were clearly the best. We would suggest the use of a prediction window of 9 residues and VTR parameters because they gave slightly better correlation on a test set of 38 proteins, because they separate flexible regions more clearly on plots, and because they gave much better correlation when used in flexibility indices. It seems that the accuracy of the sliding window technique is approaching its limit and it might be difficult to improve it significantly. The same sort of limits have also been met in secondary structural predictions where the average accuracy has been for a decade about the same in spite of numerous new methods.[71,72] These limits in prediction techniques are presumably due to intrinsic limitations of the statistical methods, which cannot take into account all the different features of complicated protein structures. Somewhat improved predictions might be obtained with neural nets and other knowledge based systems.

Flexibility profiles can be useful in several ways. When joined with sequence analysis and structural predictions, they can add to our understanding of proteins. In addition to being used for epitope searches, flexibility calculations can be applied in studies concerning sequence and structural similarity, molecular modeling, and protein engineering.

ACKNOWLEDGMENTS

Drs. Lennart Nilsson and Adrian Goldman are thanked for critical reading of the manuscript.

REFERENCES

1. Käiväräinen, A. I. "Solvent-Dependent Flexibility of Proteins and Principles of Their Function." Reidel, Dordrecht, 1985.
2. Vihinen, M. Relationship of protein flexibility to thermostability. Prot. Eng. 1:477–480, 1987.
3. Fontana, A. Structure and stability of thermophilic enzymes. Studies on thermolysin. Biophys. Chem. 29:181–193, 1988.
4. Varley, P. G. Pain, R. H. Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and *Thermus thermophilus*. J. Mol. Biol. 220: 531–538, 1991.
5. Artymiuk, P. J., Blake, C. C. F., Grace, D. E. P., Oatley, S. J., Phillips, D. C., Sternberg, M. J. E. Crystallographic studies of the dynamic properties of lysozyme. Nature (London) 280:563–568, 1979.
6. Bennett, W. S., Steitz, T. A. Glucose-induced conformational change in yeast hexokinase. Proc. Natl. Acad. Sci. U.S.A. 75:4848–4852, 1978.
7. Farnum, M. F., Magde, D., Howell, E .E., Hirai, J. T., Warren, M. S., Grimsley, J. K., Kraut, J. Analysis of hydride transfer and cofactor fluorescence decay in mutants of dihydrofolate reductase: possible evidence for participation of enzyme molecular motions in catalysis. Biochemistry 30:11567–11579, 1991.
8. Huston, E. E., Grammer, J. C., Yount, R. G. Flexibility of myosin heavy chain: direct evidence that the region containing SH1 and SH2 can move 10 Å under influence of nucleotide binding. Biochemistry 27:8945–8952, 1988.
9. Novotny, J., Handschumacher, H., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A., Rose, G. D. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc. Natl. Acad. Sci. U.S.A. 83:226–230, 1986.
10. Tanaka, T., Kato, H., Nishioka, T., Oda, J. Mutational and proteolytic studies of a flexible loop in glutathione synthetase from *Escherichia coli* B: The loop and arginine 233 are critical for the catalytic reaction. Biochemistry 31: 2259–2265, 1992.
11. Barford, D., Johnson, L. N. The allosteric transition of glycogen phosphorylase. Nature (London) 340:609–616, 1989.
12. Karplus, P. A., Schulz, G. E. Prediction of chain flexibility in proteins, Naturwissenschaften 72:212–213, 1985.
13. Bhaskaran, R., Ponnuswamy, P.K. Positional flexibilities of amino acid residues in globular proteins. Int. J. Peptide Prot. Res. 32:241–255, 1988.
14. Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A. M., Colonna, G. Flexibility plot of proteins. Prot. Eng. 2:497–504, 1989.
15. Enomaa, N., Heiskanen, T., Halila, R., Sormunen, R., Seppälä, R., Vihinen, M., Peltonen, L. Human aspartylglucosaminidase. A biochemical and immuno-cytochemical characterization of the enzyme in normal and aspartylglucosaminuria fibroblasts. Biochem. J. 286:613–618, 1992.
16. Parker, J. M. R., Guo, D., Hodges, R. S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 25:5425–5432, 1986.
17. Hopp, T. P. Use of hydrophilicity plotting procedures to identify protein antigenic segments and other interaction sites. Methods Enzymol. 178:571–585, 1989.
18. Krchnak, V., Mach, O., Maly, A. Computer prediction of potential immunogenic determinants from protein amino acid sequence. Anal. Biochem. 165:200–207, 1987.
19. Jameson, B. A., Wolf, H. The antigenic index: A novel algorithm for predicting antigenic determinants. Comput. Appl. Biosci. 4:181–186, 1988.
20. Stern, P. S. Predicting antigenic sites in proteins. Trends Biotech. 9:163–169, 1991.
21. Boberg, J., Salakoski, T., Vihinen, M. Selection of a representative set of structures from Brookhaven Protein Data Bank. Proteins 14:265–276, 1992.
22. Menendez-Arias, L., Argos, P. Engineering protein thermal stability. Sequence statistics point to residue substitutions in α-helices. J. Mol. Biol. 206:397–406, 1989.
23. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387–395, 1984.
24. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:537–542, 1977.
25. Claverie, J.-M., Daulmerie, C. Smoothing profiles with sliding windows: better to wear a hat! CABIOS 7:113–115, 1991.
26. von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol. 225:487–494, 1992.
27. Vihinen, M., Euranto, A., Luostarinen, P., Nevalainen, O. MULTICOMP, a program package for multiple sequence comparison. CABIOS 8:35–38, 1992.
28. Vihinen, M., Simultaneous comparison of several sequences. Methods Enzymol. 183:447–456, 1990.
29. Westhof, E., Altschuh, D., Moras, D., Bloomer, A. C., Mondragon, A., Klug, A., Van Regenmortel, M. H. V. Correlation between segmental mobility and the location of antigenic determinants in proteins. Nature (London) 311:123–126, 1984.
30. Atassi, M. Z. Antigenic structures of proteins. Their determination has revealed important aspects of immune recognition and generated strategies for synthetic mimicking of protein binding sites. Eur. J. Biochem. 145:1–20, 1984.
31. Dyall-Smith, M. L., Lazdins, I., Tregear, G. W., Holmes, I. H. Location of major antigenic sites involved in rotavirus serotype-specific neutralization. Proc. Natl. Acad. Sci. U.S.A. 83:3465–3468, 1986.
32. Argos, P., Fuller, S. D. A model for the hepatitis B virus core protein: prediction of antigenic sites and relationship to RNA virus capsid proteins. EMBO J. 7:819–824, 1988.
33. Hopp, T. P., Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78:3824–3828, 1981.
34. Vihinen, M. and Torkkila, E., HYDRO, a program for protein hydropathy prediction. Comput. Meth. Progr. Biomed. 41:121–129, 1993.
35. Kuroda, S., Tanizawa, K., Sakamoto, Y., Tanaka, H., Soda, K. Alanine dehydrogenases from two *Bacillus* species with distinct thermostabilities: molecular cloning, DNA and protein sequence determination, and structural comparison with other NAD(P)$^+$-dependent dehydrogenases. Biochemistry 29:1009–1015, 1990.
36. Garcia-Conzalez, M. D., Martin, J. F., Vigal, T., Liras, P. Characterization, expression in *Streptomyces lividans*, and processing of the amylase of *Streptomyces griseus* IMRU 3570: two different amylases are derived from the same gene by an intracellular processing mechanism. J. Bacteriol. 173:2451–2458, 1991.

37. Vigal, T., Gil, J. A., Daza, A., Garcia-Gonzalez, M. D., Martin, J. F. Cloning, characterization and expression of an α-amylase gene from *Streptomyces griseus* IMRU3570. Mol. Gen. Genet. 225:278–288, 1991.
38. Siggens, K. W. Molecular cloning and characterization of the beta-amylase gene from *Bacillus circulans*. Mol. Microbiol. 1:86–91, 1987.
39. Kitamoto, N., Yamagata, H., Kato, T., Tsukagoshi, N., Udaka, S. Cloning and sequencing of the gene encoding thermophilic β-amylase of *Clostridium thermosulfurogenes*. J. Bacteriol. 170:5848–5854, 1988.
40. Borriss, R., Olsen, O., Thomsen, K. K., von Wettstein, D. Hybrid *Bacillus* endo-(1-2,1-4)-β-glucanases: construction of recombinant genes and molecular properties of the gene products. Carlsberg Res. Commun. 54:41–54, 1989.
41. Nakamura, N., Horikoshi, K. Purification and properties of cyclodextrin glycosyltransferase of an alkalophilic *Bacillus* sp., Agric. Biol. Chem. 40:935–941, 1976.
42. Kaneko, T., Hamamoto, T., Horikoshi, K. Molecular cloning and nucleotide sequence of the cyclomaltodextrin glucanotransferase gene from the alkalophilic *Bacillus* sp. strain no. 38-2. J. Gen. Microbiol. 134:97–105, 1988.
43. Yamashita, I., Itoh, T., Fukui, S. Cloning and expression of the *Saccharomycopsis fibuligera* glucoamylase gene in *Saccharomyces cerevisiae*. Appl. Microbiol. Biotechnol. 23: 130–133, 1985.
44. Dohmen, R. J., Strasser, A. W. M., Dahlems, U. M., Hollenberg, C. P. Cloning of *Schwanniomyces occidentalis* glucoamylase gene (GAM1) and its expression in *Saccharomyces cerevisiae*. Gene 95:111–121, 1990.
45. Nunberg, J. H., Meade, J. H., Cole, G., Lawyer, F. C., McCabe, P., Schweickart, V., Tal, R., Wittman, V. P., Flatgaard, J. E., Innis, M. A. Molecular cloning and characterization of the glucoamylase gene of *Aspergillus awamori*. Mol. Cell. Biol. 4:2306–2315, 1984.
46. Evans, R., Ford, C., Sierks, M., Nikolov, Z., Svensson, B. Activity and thermal stability of genetically truncated forms of *Aspergillus* glucoamylase. Gene 91:131–134, 1990.
47. Schlatter, D., Kriech, O., Suter, F., Zuber, H. The primary structure of the psychrophilic lactate dehydrogenase from *Bacillus psychrosaccharolyticus*. Biol. Chem. Hoppe-Seyler 368:1435–1446, 1987.
48. Zülli, F., Schneiter, R., Urfer, R., Zuber, H. Engineering thermostability and activity of lactate dehydrogenases from *Bacilli*. Biol. Chem. Hoppe-Seyler 372:363–372, 1991.
49. Stangl, D., Widerkehr, F., Suter, F., Zuber, H. The complete amino-acid sequence of the mesophilic L-lactate dehydrogenase from *Bacillus megaterium*. Biol. Chem. Hoppe-Seyler 368:1157–1166, 1987.
50. Zülli, F., Weber, H., Zuber, H. Nucleotide sequences of lactate dehydrogenase genes from the thermophilic bacteria *Bacillus stearothermophilus, B. caldolyticus* and *B. caldotenax*. Biol. Chem. Hoppe-Seyler 368:1167–1177, 1987.
51. Zülli, F., Weber, H., Zuber, H. Analysis of structural elements responsible for the differences in thermostability and activation by fructose 1,6-bisphosphate in lactate dehydrogenases from *B. stearothermophilus* and *B. caldolyticus* by protein engineering. Biol. Chem. Hoppe-Seyler 371: 655–662, 1990.
52. Kunai, K., Machida, M., Matsuzawa, H., Ohta, T. Nucleotide sequence and characteristics of the gene for L-lactate dehydrogenase of *Thermus caldophilus* GK24 and the deduced amino-acid sequence of the enzyme. Eur. J. Biochem. 160:433–440, 1986.
53. Vasantha, N., Thompson, L. D., Rhodes, C., Banner, C., Nagle, J., Filpula, D. Genes for alkaline protease and neutral protease from *Bacillus amyloliquefaciens* contain a large open reading frame between the regions coding for signal sequence and mature protein. J. Bacteriol. 159:811–819, 1984.
54. Yang, M. Y., Ferrari, E., Henner, D. J. Cloning of the neutral protease gene of *Bacillus subtilis* and the use of the cloned gene to create in vitro-derived deletion mutants. J. Bacteriol. 160:15–21, 1984.
55. Sidler, W., Kumpf, B., Peterhans, B., Zuber, H. A neutral proteinase produced by *Bacillus cereus* with high sequence homology to thermolysin: Production, isolation and characterization. Appl. Microbiol. Biotechnol. 25:18–24, 1986.
56. Sidler, R., Niederer, E., Suter, F., Zuber, H. The primary structure of *Bacillus cereus* neutral proteinase and comparison with thermolysin and *Bacillus subtilis* neutral proteinase. Biol. Chem. Hoppe-Seyler 367:643–657, 1986.
57. Takagi, M., Imanaka, T., Aiba, S. Nucleotide sequence and promoter region for the neutral protease gene from *Bacillus stearothermophilus*. J. Bacteriol. 163:824–831, 1985.
58. Imanaka, T., Shibazaki, M., Takagi, M. A new way of enhancing the thermostability of proteases. Nature (London) 324:695–697, 1986.
59. van den Burg, B., Enequist, H. G., van der Haar, M. E., Eijsink, V. G. H., Stulp, B. K., Venema, G. A highly thermostable neutral protease from *Bacillus caldolyticus*: Cloning and expression of the gene in *Bacillus subtilis* and characterization of the gene product. J. Bacteriol. 173: 4107–4115, 1991.
60. Titani, K., Hermodson, M. A., Ericsson, L. H., Walsch, K. A., Neurath, H. Amino-acid sequence of thermolysin. Nature (London) 238:35–37, 1972.
61. Kubo, M., Imanaka, T. Cloning and nucleotide sequence of the highly thermostable neutral protease from *Bacillus stearothermophilus*. J. Gen. Microbiol. 134:1883–1892, 1988.
62. Kubo, M., Murayama, K., Seto, K., Imanaka, T. Highly thermostable neutral protease from *Bacillus stearothermophilus*. J. Ferment. Technol. 66:13–17, 1988.
63. Hilzeman, R. A., Hagie, F. E., Hayflick, J. A., Chen, C. Y., Seeburg, P. H., Derynck, R. The primary structure of the *Saccharomyces cerevisiae* gene for phosphoglycerate kinase. Nucleic Acids Res. 10:7791–7808, 1982.
64. Bowen, D., Littlechild, J. A., Fothergill, J. E., Watson, H. C., Hall, L. Nucleotide sequence of the phosphoglycerate kinase gene from the extreme thermophile *Thermus thermophilus*. Biochem. J. 254:509–517, 1988.
65. Wells, J. A., Ferrari, E., Henner, D. J., Estell, D. A., Chen, E. Y. Cloning, sequencing and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis*. Nucleic Acids Res. 11:7911–7925, 1983.
66. Wells, J. A., Powers, D. B. *In vivo* formation and stability of engineered disulfide bonds in subtilisin. J. Biol. Chem. 261:6564–6570, 1986.
67. Kleine, R. Properties of thermitase, a thermostable serine protease from *Thermoactinomyces vulgaris*. A. Biol. Med. Chem. Germ. 41:89–102, 1982.
68. Meloun, B., Baudys, M., Kostka, V., Hansdorf, G., Frömmel, C., Höhne, W. E. Complete primary structure of thermitase from *Thermoactinomyces vulgaris* and its structural features related to the subtilisin-type proteinases. FEBS Lett. 183:195–199, 1985.
69. Kwon, S.-T., Terada, I., Matsuzawa, T., Ohta, T. Nucleotide sequence of the gene for aqualysin I (a thermophilic alkaline serine protease) of *Thermus aquaticus* YT-1 and characteristics of the deduced primary structure of the enzyme. Eur. J. Biochem. 173:491–497, 1988.
70. Matsuzawa, H., Tokugawa, K., Hamaoki, M., Mizoguchi, M., Taguchi, H., Terada, I., Kwon, S.-T., Ohta, T. Purification and characterization of aqualysin I (a thermophilic alkaline serine protease) produced by *Thermus aquaticus* YT-1. Eur. J. Biochem. 171:441–447, 1988.
71. Kabsch, W., Sander, C. How good are predictions of protein secondary structure? FEBS Lett. 155:179–182, 1983.
72. Rost, B., Schneider, R., Sander, C. Progress in protein structure prediction? Trends Biochem. Sci. 18:120–123, 1993.