2306–2312 *Nucleic Acids Research, 1998, Vol. 26, No. 10*

Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites

Yael Mandel-Gutfreund and Hanah Margalit*

Department of Molecular Genetics and Biotechnology, The Hebrew University–Hadassah Medical School, PO Box 12272, Jerusalem 91120, Israel

Received February 17, 1998; Accepted March 21, 1998

ABSTRACT

Inspection of the amino acid–base interactions in protein–DNA complexes is essential to the understanding of specific recognition of DNA target sites by regulatory proteins. The accumulation of information on protein–DNA co-crystals challenges the derivation of quantitative parameters for amino acid–base interaction based on these data. Here we use the coordinates of 53 solved protein–DNA complexes to extract all non-homologous pairs of amino acid–base that are in close contact, including hydrogen bonds and hydrophobic interactions. By comparing the frequency distribution of the different pairs to a theoretical distribution and calculating the log odds, a quantitative measure that expresses the likelihood of interaction for each pair of amino acid–base could be extracted. A score that reflects the compatibility between a protein and its DNA target can be calculated by summing up the individual measures of the pairs of amino acid–base involved in the complex, assuming additivity in their contributions to binding. This score enables ranking of different DNA binding sites given a protein binding site and *vice versa* and can be used in molecular design protocols. We demonstrate its validity by comparing the predictions using this score with experimental binding results of sequence variants of zif268 zinc fingers and their DNA binding sites.

INTRODUCTION

Recent molecular and structural studies reinforce our understanding of the stereochemical principles that guide specific recognition of DNA by proteins. Structural complementarity between the protein and DNA binding sites and compatibility between the interacting groups in the protein side chains and DNA base edges are the principal determinants of specificity. The crucial role of the latter has been demonstrated both in solved crystals of protein–DNA complexes and in binding experiments of combinatorial libraries of DNA and protein binding elements (see for example 1–4). Both the structural and molecular approaches have shown that the amino acid–base interactions in the complexes are achieved mainly by hydrogen bonds and by hydrophobic interactions. Furthermore, many of the hydrogen bonds comply with the hydrogen bonding potential of the partners involved, as proposed by Seeman *et al.* (5).

Can hydrogen bonding and hydrophobic considerations by themselves serve for delineating guidelines that will enable prediction of favorable DNA binding sites given a protein binding site, and *vice versa*? The experimental data suggest that this is not the case. There are examples where preferences beyond what would have been expected from the hydrogen bonding potential of the participating residues are observed. For example, in recent compilations of all interactions that were identified in crystallographically solved protein–DNA complexes it was observed that lysine favors interactions with guanine over adenine and that aspartic acid and glutamic acid interact almost solely with cytosine (6,7). These preferences may be due to electrostatic attraction between the charged side chains of specific amino acids and the overall net charge of a particular base in the DNA groove. Other preferences that cannot be explained just by the hydrogen bonding potential of the residues were also observed in experiments with variant protein and DNA sequences. For example, in their binding experiments using sequence variants of the zif268 second zinc finger and libraries of DNA triplets Choo and Klug (3) found that histidine in the second position of the zinc finger exclusively favored guanine. This preference cannot be explained by hydrogen bond considerations, since histidine in the crystal structure of the zif268–DNA complex interacts with guanine through its hydrogen bond acceptor in position N7 and in principle this interaction could be fulfilled also by adenine, which contains an identical atom in that position. These and other examples suggest that a general recognition code, based on theoretical considerations only, may be unattainable.

An alternative approach would be to extract knowledge-based principles from the data accumulated in solved protein–DNA co-crystals and from binding experiments of sequence variants. Indeed, these two directions have recently been exploited in attempts to derive quantitative parameters for amino acid–base interactions. Suzuki and Yagi (8) used a heuristic approach to assign scores to pairs of amino acid–base, relying on the chemical nature of the participants and on the preferences found in solved protein–DNA complexes. Based on this approach quite a few amino acid–base combinations were given the same scores, resulting in insufficient discrimination between different protein–DNA binding sites. Lustig and Jernigan (9) derived a measure for amino acid–base interaction energies on the basis of relative base preferences for given amino acids, extracted from binding experiments of sequence

*To whom correspondence should be addressed. Tel: +972 2 6758614; Fax: +972 2 6784010; Email: hanah@md2.huji.ac.il

variants. Their quantitative measures were limited to the pairs of amino acid–base tested in the experiments they relied on (4). They suggested that these energies can characterize the most important interactions of bases and amino acids.

The continuously increasing number of crystallographically solved protein–DNA complexes challenges the derivation of a quantitative measure for all possible amino acid–base interactions from their frequencies in the three-dimensional structures of the complexes, similarly to the extraction of knowledge-based amino acid–amino acid contact energies. Pairwise contact potentials for amino acid–amino acid interactions were derived empirically from protein tertiary structures by several groups and were found to be very useful in fold recognition schemes (reviewed in 10). The underlying assumption there was that in a sufficiently large sample of protein structures the number of spatially close pairs of amino acids reflects the average likelihood of interaction between the two types of amino acids involved. The approach, therefore, was to count the number of side chain contacts between a given pair of amino acids and to extract a pairwise contact potential using Boltzmann's formalism. These empirical potentials were used to evaluate sequence–structure fit, i.e. to select for a given protein sequence the most compatible three-dimensional fold from a library of known structures, and *vice versa*. In the present study we describe a similar analysis based on the frequency of pairs of amino acid–base that are involved in specific interactions in the solved protein–DNA complexes. A quantitative measure for base–amino acid interaction is obtained by computing the log odds of the observed pair frequencies and those expected at random. Although the total number of amino acid–base interactions in solved complexes is significantly smaller than the number of amino acid–amino acid interactions in solved protein structures, still the measure obtained seems to reflect the likelihood of interaction of a given pair of amino acid–base. This is supported by the correspondence between the computed scores using this measure and results of binding experiments.

Recently, new approaches to study sequence-specific DNA recognition have been introduced by several groups. These approaches involve screening of DNA libraries for binding by given sequence variants of the zinc fingers of the transcription factor zif268, and *vice versa* (2,3,11–13). zif268, which belongs to the Cys_2His_2 family of zinc finger proteins, provides a convenient system to study the specificity of recognition between an amino acid and a DNA base. This is due to the simplicity of its structure and mode of interaction with DNA. zif268 contains three zinc fingers which bind in a modular fashion to an array of three DNA triplets; each finger binds a DNA triplet (14,15). A consensus binding pattern for the three fingers was inferred based on the crystal structures of the complex and on binding studies (16). It involves three amino acids, one preceding the α-helix of the zinc finger and two that are included in it, recognizing specifically the DNA triplet in a one-to-one manner, as illustrated in Figure 1. Based on this simple binding pattern, preferences for individual pairs of amino acid–base could be evaluated experimentally by testing the binding of substituted sequences of the protein and DNA binding elements. We used these experimental data to judge the feasibility of the quantitative measure and show that the computed scores succeed fairly well in predicting the hierarchy of binding and in distinguishing between good and poor binding sites. The agreement between the computational results and the experimental data suggests that these knowledge-based quantitative parameters can be used for prediction of potential binding sites in molecular design experiments.

MATERIALS AND METHODS

Database of crystal structures

Our data set contains 53 crystallographically solved protein–DNA complexes. Forty six were extracted from the PDB database (17): 1aay, 1apl, 1ber, 1bhm, 2bop, 1cdw, 1cma, 3cro, 1d66, 2dgc, 2drp, 1eri, 1fjl, 1fos, 1glu, 1hcq, 1hcr, 1hdd, 1ign, 1ihf, 1lat, 1mey, 1mdy, 1nfk, 2nll, 1oct, 2or1, 1par, 1pdn, 1per, 1pnr, 1pri, 1pue, 1pyi, 1rpe, 1rva, 1srs, 1tro, 1trr, 1tup, 1yrn, 1ysa, 1ytb, 1ubd. The other six structures were extracted from the NDB database (18): pdt009, pdt013, pdt020, pdt022, pdt023, pdt027, pdt031.

Determination of amino acid–base contacts

All pairs of amino acid–DNA base which are in contact through the amino acid side chain and the DNA base edge, either by hydrogen bonds or by hydrophobic interactions, were extracted from the atomic coordinates of the complexes. Homologous interactions were excluded, resulting in a non-redundant data set of pairs of amino acid–base. In the current study only interactions that involve atoms in the DNA major groove were included. Determination of the amino acid side chain and DNA base edge atoms that can participate in hydrogen bonds was based on Ippolito *et al.* (19) and Seeman *et al.* (5), respectively. Hydrogen bonds were defined as in Mandel-Gutfreund *et al.* (7). Hydrophobic interactions were determined as carbon–carbon interactions within a distance ≤ 4.0 Å, involving the methyl group of thymine or the C5 group of cytosine and the carbons in the side chains of the hydrophobic/aromatic amino acids. Carbons which are covalently bound to polar atoms in the aromatic side chains were excluded.

Determination of a scoring measure for pairs of amino acid–base

The data of non-homologous pairs of amino acid–base that are in contact were arranged in a 'frequency matrix', rows and columns

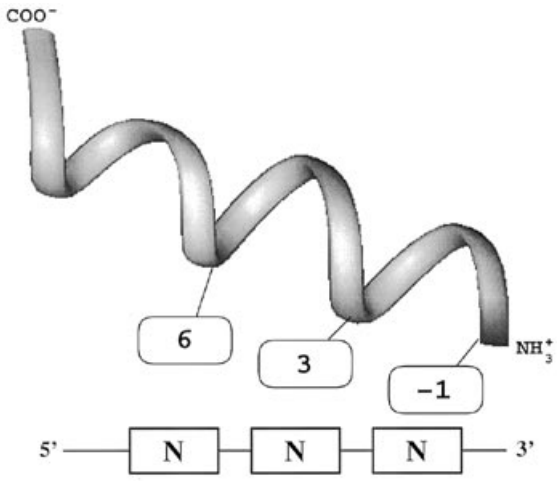

Figure 1. A schematic model of the Cys_2His_2 zinc finger–DNA consensus binding pattern, based on the specific interactions observed in the crystal structure of the zif268–DNA complex and binding studies (16). As illustrated, three amino acids of the zinc finger, in positions –1, 3 and 6 with respect to the α-helix, contact three adjacent bases on one DNA strand in an 'anti-parallel' manner.

representing amino acids and bases, respectively. The frequency distribution of the pairs was compared with that expected at random, based on general frequencies of amino acids and bases. The SWISSPROT database was used to extract the general frequencies of amino acids in known proteins. As for the base frequencies, since the DNA target sites in our data were from different organisms, there was no one database that we could use, thus an equal probability of 0.25 for all four bases was used. The expected frequency of a pair of amino acid–base was obtained as the product of the two appropriate random frequencies. A quantitative measure for amino acid–base interaction was obtained by calculating the log odds (log likelihood ratio) for each pair:

$$S_{ij} = \ln[f_{ij}/(f_i \times f_j)]$$

where f_{ij} is the pair frequency of a specific amino acid i and base j, f_i is the frequency of amino acid i (i = 1,20) and f_j is 0.25 (j = 1,4). When the number of pairs of a certain type was equal to zero, so that $\ln(f_{ij})$ could not be defined, two approaches were applied, according to the source of the zero frequency. Pairs which are impossible because they lack complementary chemical groups that can be involved in a direct interaction (annotated NA in Table 1) were scored as (–3.93), the lowest possible score in the table. For pairs which are theoretically possible but did not occur in the solved complexes we arbitrarily increased their count to 0.1 and their scores were calculated as for the other cases. The total number of all pairs was increased accordingly.

The score of an interaction between protein and DNA binding elements is obtained by summation of the individual scores of the interacting pairs of amino acid–base in the complex, assuming that the contributions of individual pairs are independent of one another.

RESULTS

To evaluate the likelihood of an amino acid and a DNA base to interact, non-homologous specific amino acid–base interactions were extracted from a data set of 53 crystallographically solved protein–DNA complexes, including transcription factors and restriction enzymes (listed in Materials and Methods). The current analysis is focused on interactions that involve atoms in the major groove only, since these constitute most of the specific interactions in the solved structures and since the pattern of donors and acceptors of these atoms is unique for each DNA base and is considered to be the main contributor to specific recognition of the bases by the different amino acids. Three hundred non-homologous contacts were observed, including hydrogen bonds and hydrophobic interactions between amino acid side chains and DNA major groove atoms. In the analysis each pair of amino acid–base was considered only once, independent of the number of hydrogen bonds and/or hydrophobic interactions that are involved in forming that pair. This resulted in a total of 218 different pairs used in the following steps of the analysis. The occurrences of the 20 × 4 pairs of amino acid–base are summarized in Table 1. The frequency distribution of the different pairs is very similar to that observed by us previously based on a smaller data set (7). As predicted by Seeman *et al.* (5) and observed in many other studies (3,6,7,20), the most frequent pair is Arg–G. Glutamine and aspargine interact preferably with adenine and the two negatively charged amino acids glutamic acid and aspartic acid interact almost solely with cytosine. The present analysis includes also hydrophobic interactions, involving interactions between the major groove carbon atoms of thymine (C5M) and cytosine (C5) and the carbons of the hydrophobic and aromatic amino acids. Only interactions that involve the methyl group of thymine were observed and these were formed relatively frequently with alanine and isoleucine.

Table 1. Observed frequency of 20 × 4 pairs of amino acid–DNA base

	G	A	T	C	Total
Gly	NA	NA	NA	NA	0
Ala	NA	NA	8	0	8
Val	NA	NA	3	0	3
Ile	NA	NA	6	0	6
Leu	NA	NA	2	0	2
Phe	NA	NA	1	2	3
Trp	0	NA	0	NA	0
Tyr	0	0	3	2	5
Met	0	1	2	1	4
Cys	0	1	0	1	2
Thr	0	3	3	1	7
Ser	6	2	3	2	13
Gln	2	7	3	0	12
Asn	4	17	5	5	31
Glu	NA	1	NA	6	7
Asp	NA	0	NA	8	8
His	6	2	3	1	12
Arg	44	4	10	NA	58
Lys	28	3	4	NA	35
Pro	NA	NA	2	0	2
Total	90	41	58	29	218

Table 2. Scoring matrix for 20 × 4 pairs of amino acid–DNA base

	G	A	T	C
Gly	–3.93	–3.93	–3.93	–3.93
Ala	–3.93	–3.93	0.66	–3.72
Val	–3.93	–3.93	–0.17	–3.57
Ile	–3.93	–3.93	0.65	–3.44
Leu	–3.93	–3.93	–0.94	–3.93
Phe	–3.93	–3.93	–0.81	–0.12
Trp	–1.96	–3.93	–1.96	–3.93
Tyr	–2.87	–2.87	0.54	0.13
Met	–2.58	–0.28	0.42	–0.28
Cys	–2.23	0.07	–2.23	0.07
Thr	–3.46	–0.06	–0.06	–1.16
Ser	0.42	–0.68	–0.28	–0.68
Gln	–0.09	1.16	0.31	–3.09
Asn	0.48	1.93	0.71	0.71
Glu	–3.93	–1.24	–3.93	0.55
Asp	–3.93	–3.37	–3.93	1.01
His	1.56	0.46	0.87	–0.23
Arg	2.74	0.34	1.25	–3.93
Lys	2.16	–0.08	0.21	–3.93
Pro	–3.93	–3.93	–0.30	–3.29

Scores were calculated using the formula: $\ln[f_{ij}/(f_i \times 0.25)]$, where f_{ij} is the frequency of the pair between amino acid (i) and DNA base (j). f_i is the frequency of amino acid i in the SWISSPROT database of protein sequences and 0.25 is the equal probability assumed for each of the DNA bases.

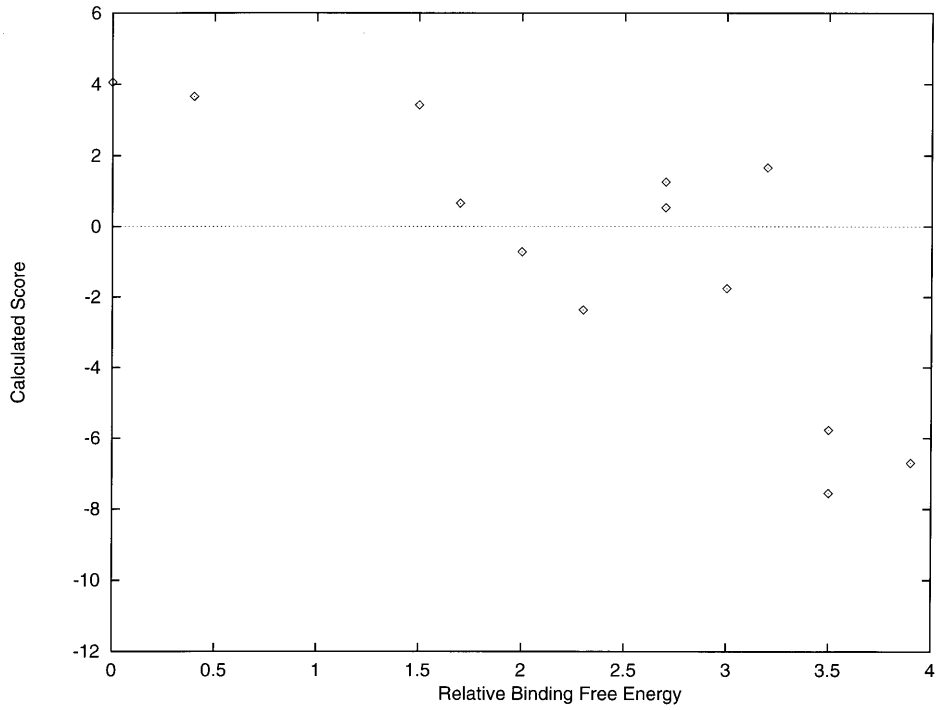

Figure 2. Correlation between the calculated scores and experimentally determined binding free energies (in kcal/mol) for 13 different DNA triplets bound to a zinc finger variant containing the residues QDR in positions –1, 3 and 6 respectively (4). The Spearman rank correlation coefficient is –0.79.

The frequency table of pairs of amino acid–base was used to generate quantitative measures for amino acid–base interactions. The likelihood ratio for each pair of amino acid–base was defined as the ratio between the frequency of the specific pair in the protein–DNA complexes and the theoretical probability of obtaining such a pair, based on the overall frequencies of the amino acids in all known proteins and on an equal probability of 0.25 for the DNA bases (see Materials and Methods and Discussion). Table 2 lists the quantitative measures for all pairs of amino acid–base observed in the solved complexes, obtained by calculating the natural logarithm of the likelihood ratios. The score obtained by summing up these measures for all pairs involved in a complex is expected to reflect the compatibility between the respective DNA and protein binding sites.

To assess the validity of the computed scores we compared the computational results with experimental binding data of sequence variants of zif268 zinc fingers and their DNA binding sites. For each combination of an amino acid triplet with a DNA triplet the pairs of amino acid–base were determined according to the binding model (Fig. 1) and a score was obtained by summing up the quantitative measures (Table 2) for all three pairs involved. The inherent assumption is that no changes in the protein–DNA interface occur upon substitution and that the same positions in the protein and DNA stay in contact (2–4,11–13). Such a fixed binding framework may result from the orientation of the DNA binding domain of the zinc finger relative to the DNA and the spacing between the amino acids used for contacting the DNA (16). As demonstrated in Figure 2, a significant correlation (r = –0.79, $P < 0.001$) was obtained between the calculated scores and relative free energies extracted by Desjarlais and Berg (4) from their experimentally determined dissociation constants. A lower correlation coefficient (r = –0.49), although still statistically significant ($P < 0.025$), was obtained when comparing the calculated scores with another set of experimentally determined dissociation constants (3). Yet while the computed scores in the latter predicted the hierarchy of binding less accurately compared with that observed experimentally, they could be used successfully to distinguish between binding and non-binding triplets of protein–DNA complexes. Since data on experimentally determined binding constants are scarce, such tests of the scoring scheme should be performed again when more experimental data are available.

Another kind of experimental data was provided by selection studies, where either the most favorable DNA triplet was selected for a given protein binding site among all possible triplets, or the most favorable amino acid combination was selected for a given DNA triplet among many sequences of amino acids (2–4,12,13). Ranking of a combination by the computed scores can be made either among all 64 possible DNA triplets or among all 8000 possible amino acid triplets, or even among all 512 000 possible combinations, depending on the experiment used to assess the computations. For example, Figure 3 demonstrates the computed scores for all 64 DNA triplet combinations bound to a variant of finger 2 of Zif268, **R**GD**A**LT**S**HER, where only the relevant positions of the recognition helix, –1, 3 and 6 (noted in bold), were taken into consideration. As can be seen, the two DNA triplets GTG and TTG which were selected experimentally (3) were ranked in the two highest ranks, with scores of 3.82 and 3.12, respectively. Such comparisons were carried out for all the results of selection experiments documented by Choo and Klug (3) and by Desjarlais and Berg (4). For each given triplet of interacting amino acids the binding scores with all possible 64 DNA combinations were computed and ranked. The rank of the experimentally selected DNA triplet among the 64 triplets was documented. Figure 4 summarizes the ranking, based on the

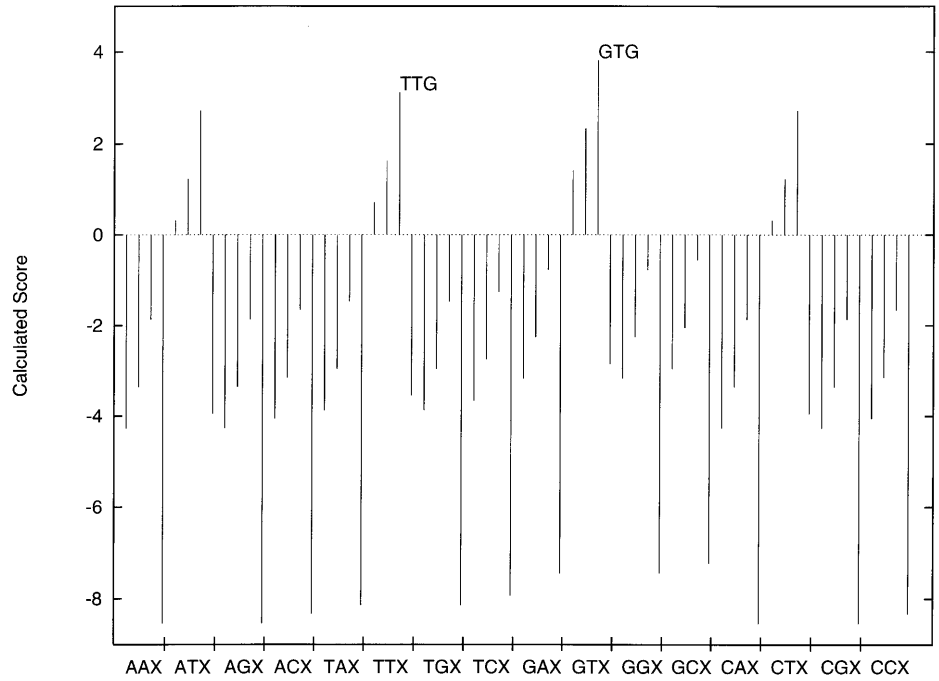

Figure 3. Predicted scores for binding of a zinc finger variant **R**GD**A**LT**S**HER (bold capital letters indicate amino acids involved in contacts) to all possible 64 DNA triplets. The DNA triplets are ordered in the figure as indicated, where X changes in the order A, T, G, C. The triplets GTG and TTG are the ones selected experimentally (3).

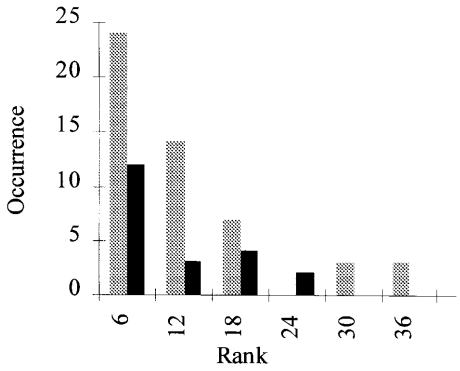

Figure 4. Summary of the ranks obtained by the computed scores for experimentally selected DNA binding sites by given zinc finger variants (see text), based on data from 52 selection studies by Choo and Klug (3) (gray) and 21 selection studies by Desjarlais and Berg (4) (black).

computed scores, for a total of 73 experimentally selected DNA binding sites by zif268 finger 2 variants (data from 3,4). The rankings are given in absolute numbers (1–64) and the height of the histogram represents the number of experimentally selected triplets ranked in that range by the computed score. It is noteworthy that ~50% of the experimentally selected protein–DNA pairs from the two different data sets give scores ranked among the highest six triplets out of all possible DNA triplets. Only three experimentally selected triplets were ranked below the 32nd rank.

The computed scores were also interpreted in the other direction, given a DNA triplet and selecting the optimal protein triplet. Table 3 summarizes the computed scores for two sets of sequences from selection studies of the substituted zif268 finger 1 recognition helix, given different DNA binding sites (sequences in sections A and B are taken from 13, table 4 and 12, table 2, respectively). Here the ranking is documented in percentiles, either among all possible 8000 amino acid triplets or among all possible 512 000 amino acid–base combinations of triplets. Two of 12 experimentally selected combinations (GCG–RDR, row 4; GAC–DNR, row 10) were ranked by us in the highest rank. Nine of 12 and 10 of 12 were ranked in the first two percentiles when the rank was calculated among 8000 combinations and among 512 000 combinations, respectively, usually in a relatively high rank. In another set of experimentally selected Zif268 finger 2 variants by given DNA triplets (data from 2) 78% of the amino acid combinations were ranked by the computed scores in the first 10 percentiles (data not shown).

DISCUSSION

One of the main challenges in molecular biology is to understand what determines the selection of DNA target sites by a regulatory protein. Examination of the solved protein–DNA complexes shows a stereochemical complementarity between the elements involved in forming the complex. These data and results of binding experiments of sequence variants indicate that the compatibility between the interacting groups in the protein and DNA plays a major role in dictating specific recognition. In its simplest view, a protein binding site will favor DNA target sites in which there is a one-to-one compatibility between the amino acids used for interaction and the DNA bases that they contact. Understanding how this compatibility is determined is the key to discerning specific recognition, and the ability to quantify amino acid–base compatibility is the basis for any predictive algorithm that will identify favorable binding sites. In the present study we took advantage of the growing number of solved protein–DNA co-crystals and extracted a quantitative measure for amino acid–base compatibility based on the observed frequencies of pairs of amino acid–base that are in contact in the solved complexes.

Table 3. Calculated scores and ranking for selected fingers by DNA triplets [based on data from Jamieson *et al*. (13) (**A**) and Rebar and Pabo (12) (**B**)]

	DNA triplet			Amino acid triplet			Score	Percentile	Percentile
	1	2	3	–1	3	6		(rank/8000)	(rank/512000)
A	G	A	C	E	N	R	5.22	1	1
	G	C	A	Q	E	R	4.45	1	1
	G	C	G	R	E	R	6.03	1	1
				R	D	R	6.49	1	1
				G	E	R	–0.64	9	14
	G	G	G	R	H	R	7.04	1	1
	G	G	A	K	H	R	4.22	1	1
	G	T	G	E	A	R	–0.53	13	14
	G	T	T	T	A	R	3.34	2	1
B	G	A	C	D	N	R	5.68	1	1
	G	C	A	R	D	R	4.09	1	1
				Q	S	R	3.22	1	1

Amino acid triplets listed in the second column are those selected experimentally for given DNA triplets (listed in the first column). Scores for each combination of the listed amino acid–base triplets were calculated based on the matrix in Table 2. Ranking of the experimentally selected amino acid triplets by the scores out of all possible 8000 combinations of amino acid triplets is given in column 4, while ranking out of all possible 512 000 combinations of amino acid–base triplets is given in column 5.

Our approach was to count the number of all different pairs of amino acid–base in the solved complexes and to extract a measure by computing the log odds of the observed frequencies and those expected if these interactions were random. The log odds matrix for base–amino acid interactions quantifies the preferences of the pairwise interactions and can be used to evaluate compatibility between protein and DNA binding sites. Log odds matrices were generated in a variety of computational studies that attempted to derive knowledge-based parameters from a data set of sequences or structures, given a particular biological question. They have been employed in the derivation of scoring matrices for amino acid–amino acid substitutions, such as the Dayhoff matrix, used for protein sequence alignment (reviewed in 21). Information content calculations in aligned sequences and generation of a specificity matrix to define a particular functional site are based on the same concept (reviewed in 22). Scores that describe the compatibility between the different amino acids and defined structural environments were extracted similarly by Eisenberg and colleagues from a database of solved protein structures and have been used to evaluate sequence–structure fit (23,24). Also, as commented by Jones and Thornton (10), many of the derived amino acid–amino acid contact potential matrices can be considered as log odds matrices, as they encode observed distributions of residue pairs in real proteins and do not seek to measure energy. Likewise, in the present study there is no attempt to ascribe an energetic meaning to the extracted values, but merely to consider them as quantitative scores that reflect compatibility between amino acids and bases.

In all examples above, the likelihood ratios compare the probability of an event occurring under two alternative hypotheses. In the case of amino acid–base interaction we compare the frequencies of pairs that appear in solved structures with the expected frequencies if these interactions were random. The expected frequency of a pair is calculated under the assumption that there is no preference for any amino acid to interact with any base, by multiplying the expected frequencies of bases and amino acids. These latter frequencies may be defined in different ways. One approach is to use the total frequencies of amino acids and bases in the data. However, by this approach interactions that involve bases and amino acids that are frequent in protein–DNA complexes become artificially weaker. Also, by using the frequencies from the data set itself the fact that some bases and amino acids participate in protein–DNA complexes at frequencies well above average is masked. To overcome these drawbacks general frequencies of amino acids and bases were used. As shown in Table 2, the quantitative measures obtained succeed in reflecting reasonably well the pair preferences. For example, the highest measures were obtained for Arg–G, Lys–G and Asn–A. The preference for these pairs in protein–DNA recognition has been shown in many structures and their possible role in protein–DNA recognition has been suggested (7,8,20). In their scoring scheme Suzuki and Yagi (8) assigned to all of them an equal score (the highest score in their scheme). The current measures, however, indicate a hierarchy of these pairs: Arg–G > Lys–G > Asn–A. Another example is hydrophobic interactions. In the scoring scheme of Suzuki and Yagi (8) most of those interactions were scored equally. Here, Ala–T and Ile–T seem to be the most favorable among hydrophobic interactions, followed by Tyr–T and Met–T. However, because of the limited size of the data set used to derive these parameters, such conjectures should be drawn with caution. It is expected that with the accumulation of more solved complexes the accuracy of such quantitative parameters will increase. The advantage of the current computational approach is that the quantitative measures can be refined systematically when the pair frequency of amino acids and bases is updated. We have recently explored the possible role of CH···O interactions in protein–DNA recognition and concluded that inclusion of these interactions in the amino acid–base frequency matrix results in a more consistent pattern of the preferred pairs, which can be explained on the basis of electrostatic considerations (25). Thus, inclusion of these interactions in future derived log odds matrices may also improve the parameters. Nevertheless, as demonstrated above and discussed below, despite the relatively small size of the structural data set used and with all reservations taken into account, using such parameters for prediction looks promising.

The usual application of log odds matrices in the context of sequence and structure analysis is to evaluate compatibility either between two sequences or between a sequence and a structure. The approach is to sum the appropriate log odds along the alignment to a score that can be used to compare different alignments. The inherent assumption in all these studies is that the contributions of the different positions along the alignment are independent. The search for compatible protein and DNA sequences that will form the most favorable protein–DNA complex can also be viewed as an alignment problem, where the log odds matrix is used to find the best match of protein and DNA sequences among several possibilities. Additivity in the contributions of the pairs involved in the complex is assumed here also, however, in this case it has some experimental support. In at least one experimental binding study additivity was inferred based on differences in the dissociation constants of all possible sequence variants of OR1 bound to the Cro protein of phage λ (26). In other studies consistent interpretations of binding results of substituted sequences (amino acids and/or DNA bases) are possible when each pair interaction is considered as independent of the other interactions (see for example 27). Ideally, possible interdependence between binding residues should be taken into account and is expected to improve prediction. This may be feasible with a significantly larger data set of solved complexes, enabling quantitative evaluation of the mutual effects of different pairs of amino acid–base on binding.

In the current study the validity of the computed scores is demonstrated by their consistency with experimental binding data of variant sequences of zif268 zinc fingers and their DNA binding sites. In most cases the computed scores succeeded in ranking highly those combinations of amino acid and DNA base triplets that were selected experimentally, although not always in the highest rank (Fig. 4). In some instances the amino acids or bases that were selected experimentally were not those that are the most compatible according to the log odds matrix. This is probably due to the effect of other factors on binding. Note that these quantitative parameters were derived from the pair interactions in a variety of protein–DNA complexes and reflect the likelihood of interaction in general. Consequently, possible position-dependent effects specific to each binding motif are masked. For example, in the zif268-like zinc fingers steric constraints that are position dependent are probably imposed by the specific orientation of the protein binding element relative to the DNA (16). Conceivably, incorporation of position-dependent effects specific to each binding motif together with the quantitative parameters will yield better predictions (6,16). In addition, there are other factors that affect binding, such as the sequence context of the binding sites, coupled interactions, where one amino acid is assisted by another in contacting the DNA (see for example 15,28), and the structure of the DNA binding site (see for example 29–31). In the case of the zif268-like zinc fingers all these additional factors were not taken into account and yet the scores calculated by the quantitative measures gave quite satisfactory predictions. This is probably due to the simple binding framework of these proteins. In more complicated cases the other factors may carry more weight and should also be considered. This requires quantitation of the different parameters, like position-dependent effects and coupled interactions, as well as prediction of the DNA structure in the binding site. Still, for all families of transcription factors in which the framework of amino acid–base interactions is defined, quantitative parameters such as those extracted here are useful for first screening and narrowing down the number of candidate sequences.

ACKNOWLEDGEMENTS

We thank Robert Jernigan, Victor Zhurkin and Ora Schueler for helpful discussions. This study was supported by a grant to H.M. from the Israel Science Foundation, administered by the Israeli Academy of Sciences and Humanities.

REFERENCES

1 Tanikawa,J., Yasukawa,T., Enari,M., Ogata,K., Nishimura,Y., Ishii,S. and Sarai,A. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 9320–9324.
2 Choo,Y. and Klug,A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11163–11167.
3 Choo,Y. and Klug,A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11168–11172.
4 Desjarlais,J.R. and Berg,J.M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11099–11103.
5 Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) *Proc. Natl. Acad. Sci. USA*, **73**, 804–808.
6 Suzuki,M. (1994) *Structure*, **2**, 317–326.
7 Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) *J. Mol. Biol.*, **253**, 370–382.
8 Suzuki,M. and Yagi,N. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 12357–12361.
9 Lustig,B. and Jernigan,R.L. (1995) *Nucleic Acids Res.*, **23**, 4707–4711.
10 Jones,D.T. and Thornton,J.M. (1996) *Curr. Opin. Struct. Biol.*, **6**, 210–216.
11 Jamieson,A.C., Kim,S.H. and Wells,J.A. (1994) *Biochemistry*, **33**, 5689–5695.
12 Rebar,E.J. and Pabo,C.O. (1994) *Science*, **263**, 671–673.
13 Jamieson,A.C., Wang,H. and Kim,S. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 12834–12839.
14 Pavletich,N.P. and Pabo,C.O. (1991) *Science*, **252**, 809–817.
15 Elrod-Erickson,M., Rould,M.A., Nekludova,L. and Pabo,C.O. (1996) *Structure*, **4**, 1171–1180.
16 Choo,Y. and Klug,A. (1997) *Curr. Opin. Struct. Biol.*, **7**, 117–125.
17 Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1978) *Arch. Biochem. Biophys.*, **185**, 584–591.
18 Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) *Biophys. J.*, **63**, 751–759.
19 Ippolito,J.A., Alexander,R.S. and Christianson,D.W. (1990) *J. Mol. Biol.*, **215**, 457–471.
20 Pabo,C.O. and Sauer,R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.
21 Altschul,S.F. (1991) *J. Mol. Biol.*, **219**, 555–565.
22 Stormo,G.D. (1990) *Methods Enzymol.*, **183**, 211–221.
23 Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
24 Bowie,J.U., Zhang,K., Wilmanns,M. and Eisenberg,D. (1996) *Methods Enzymol.*, **266**, 598–616.
25 Mandel-Gutfreund,Y., Margalit,H., Jernigan,R.L. and Zhurkin,V.B. (1998) *J. Mol. Biol.*, **277**, 1129–1140.
26 Takeda,Y., Sarai,A. and Rivera,V.M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 439–443.
27 Deng,Q.L., Ishii,S. and Sarai,A. (1996) *Nucleic Acids Res.*, **24**, 766–774.
28 Aggarwal,A.K., Rodgers,D.W., Drottar,M., Ptashne,M. and Harrison,S.C. (1988) *Science*, **242**, 899–907.
29 Kim,J.L. and Burley,S.K. (1994) *Nature Struct. Biol.*, **1**, 638–653.
30 Parkinson,G., Wilson,C., Gunasekera,A., Ebright,Y.W., Ebright,R.E. and Berman,H.M. (1996) *J. Mol. Biol.*, **260**, 395–408.
31 Rice,P.A., Yang,S., Mizuuchi,K. and Nash,H.A. (1996) *Cell*, **87**, 1295–1306.