1995

Consistencies of individual DNA base-amino acid interactions in structures and sequences

Brooke Lustig, *San Jose State University*
R L Jernigan

Nucleic Acids Research, 1995, Vol. 23, No. 22 ***4707–4711***

Consistencies of individual DNA base–amino acid interactions in structures and sequences

Brooke Lustig and Robert L. Jernigan*

Laboratory of Mathematical Biology, Room B116, Building 12B, MSC-5677, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892-5677, USA

Received July 10, 1995; Revised and Accepted October 14, 1995

ABSTRACT

Amino acid–amino acid interaction energies have been derived from crystal structure data for a number of years. Here is reported the first derivation of normalized relative interaction from binding data for each of the four bases interacting with a specific amino acid, utilizing data from combinatorial multiplex DNA binding of zinc finger domains [Desjarlais, J. R. and Berg, J. M. (1994) *Proc. Natl. Acad. Sci. USA*, 91, 11099–11103]. The five strongest interactions are observed for lysine–guanine, lysine–thymine, arginine–guanine, aspartic acid–cytosine and asparagine–adenine. These rankings for interactions with the four bases appear to be related to base–amino acid partial charges. Also, similar normalized relative interaction energies are derived by using DNA binding data for Cro and λ repressors and the R2R3 c-Myb protein domain [Takeda, Y., Sarai, A. and Rivera, V. M. (1989) *Proc. Natl. Acad. Sci. USA*, 86, 439–443; Sarai, A. and Takeda, Y. (1989) *Proc. Natl. Acad. Sci. USA*, 86, 6513–6517; Ogata, K. *et al.* (1995) submitted]. These energies correlate well with the combinatorial multiplex energies, and the strongest cases are similar between the two sets. They also correlate well with similar relative interaction energies derived directly from frequencies of bases in the bacteriophage λ operator sequences. These results suggest that such potentials are general and that extensive combinatorial binding studies can be used to derive potential energies for DNA–protein interactions.

INTRODUCTION

The interactions between DNA and proteins are critical to gene expression and regulation, and of significant importance for possible therapeutic strategies. The effects of mutagenesis on binding to DNA have been extensively studied for several proteins in several ways. The binding constants were reported by Takeda *et al.* for Cro (1) and by Sarai and Takeda for λ repressor (2) as well as by Sarai for transcription activator c-Myb (3). These were determined for sets of exhaustive single base substitutions in the native DNA sequences based on the corresponding three-dimensional structures determined from X-ray (4,5) and NMR (6).

Desjarlais and Berg used combinatorial synthesis methods to study the effects on binding of single base substitutions for an extensive set of 18 zinc finger sequences (7). Each peptide is composed of three 28–30 residue zinc finger domains. They constructed a DNA–zinc finger protein system, based on a crystal structure (8), in which the first and third of the zinc finger domains had a fixed sequence, and the central zinc finger sequence was varied. As with the two flanking domains, the central zinc finger domain has three amino acids interacting with three base-pairs of the DNA. For each of their 18 different amino acid triplets, all possible substitutions of the three base pairs were made. This constrained system was designed to ensure that the interactions of the non- substituted bases remained intact. Length-encoding of the DNA was used in order to identify the relative preferences of particular base–amino acid interactions from the corresponding intensities observed on electrophoresis gels. Our purpose in this paper is to compare the various data and to derive normalized base–amino acid interaction energies from the combinatorial binding studies and compare them to similarly derived energies for Cro, λ repressor and c-Myb, as well as make comparisons to energies derived from the operator sequences for the two repressor proteins. The present study is structure based, in that substitutions in DNA sequence are considered at sites assumed to maintain the neighboring DNA–protein contacts in the triplet. The comparisons here will show that such interaction potentials are relatively independent of the specific DNA–protein complex, and therefore ought to have some general validity for evaluating other DNA–protein complexes.

Previously, the successful development by Miyazawa and Jernigan of statistical potentials for interacting amino acid–amino acid pairs was based on a set of globular protein crystal structures (9,10). The approach there was simply to count the number of close pairs of amino acids. These total frequencies, over large numbers of protein structures, reflect the average likelihood of interaction between the two types of amino acids, independent of sequence and other structural details. In other words, if a sufficiently large sample of protein structures is considered and the backbones are sufficiently flexible then the distribution of interactions will directly reflect their individual strengths, i.e., the frequencies of non-bonded close residue pairs will reflect their interaction energies. Thus, by forming the frequencies into equilibrium constants, and by using a Boltzmann form of the equilibrium constant, specific residue–residue type interaction

* To whom correspondence should be addressed

energies were derived, as e_{IJ}, for the interaction between an I-type and a J-type residue pair (9). The total energy for a protein can then be described in terms of these pair-wise contact potentials as

$$E = \sum_{I,J=1}^{20} e_{IJ}\, n_{IJ}$$

where I and J are indices for each of the 20 types of amino acids and n_{IJ} is the number of IJ contacts in the structure. These contact energies resemble a set of pair-wise hydrophobicities (9,11). The values have been used subsequently to reconstruct approximate total energies for individual proteins, and shown to represent the energetic effects of mutagenesis (10). The validity and utility of this approach for interactions in globular proteins are well established. Here we are going to apply similar methods to DNA–protein interactions on the basis of the manifestations of relative base preferences for given amino acids and will compare apparent interaction energies derived from several sources. In the present case, somewhat fewer assumptions are required than in the intra-molecular protein case since the assumption of independent behavior is obviously more appropriate for binding between two molecules than for the intra-molecular 'binding' case.

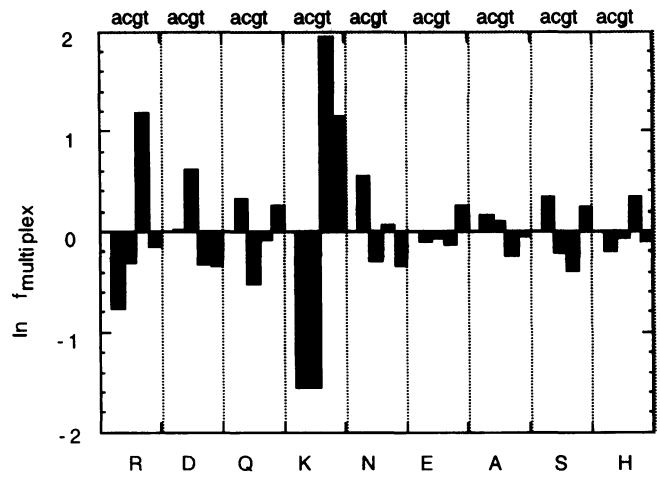

Figure 1. Logarithms of the relative frequencies, i.e. relative energies, of base–amino acid binding interactions derived from the combinatorial multiplex binding data of Desjarlais and Berg (7) for zinc finger domains. The logarithms are normalized for each set of four bases for one amino acid type. Data for the amino acid type specified on the lower abscissa is all given between the dotted vertical dividers. The order of bases is given along the top.

METHODS AND RESULTS

Relative interaction energies from multiplex binding studies

The combinatorial multiplex DNA binding studies of zinc finger protein domains include base type preferences for each of the three positions interacting with peptide triplets (7). We use those frequencies to derive relative interaction energies by first taking the logarithms of the frequencies for all occurrences of a j-type base interacting with an I-type amino acid so that the interaction energy e_{Ij} is in the form

$$e_{Ij} \sim -\ln f_{Ij} \qquad 1$$

where f_{Ij} is the sum over all 18 sets of the relative frequencies with which a base type j interacts with all occurrences of a residue type I. Some other similar data are available (12) that have not been used in this study. For each of the four bases, the relative interaction energies are then normalized as

$$\sum_{j} \ln f_{Ij} = 0 \qquad 2$$

These normalized relative interaction energies are shown in Figure 1. Although these energies are presumed to correspond to the proximate placement of an amino acid at a particular nucleotide, there remains the possibility that structures might change upon substitution. However, this assumption is strongly supported by binding experiments that show that changes in ΔG for a particular DNA–protein interaction determined from binding studies can be reliably calculated simply as the sum of individual interactions (1,2,7).

Analyses here will describe only relative changes as a result of base substitutions. We choose to do this because data currently are available for all four bases at particular positions, but not for all 20 amino acids (1–3,7). Furthermore, substitutions in the DNA are extremely unlikely to disrupt the double helix, but there is a greater possibility that the protein structure could be changed by substitutions in the amino acid sequence.

Most favorable interactions

Previously specific favorable interactions between amino acids and bases have been proposed. Several of the cases given by Seeman *et al.* (13) are also found to be strong here. From Figure 1 the five most favorable relative interaction energies are seen to involve arginine, lysine, aspartic acid and asparagine. Interactions are most favorable for guanine with arginine and lysine, and also for thymine with lysine. (The data in ref. 12 also suggest relatively strong interactions between arginine and thymine.) For those two amino acid types the most preferred pairs for interaction are guanine and thymine; whereas for aspartic acid, adenine and cytosine are the preferred pairs for interaction. This suggests the possible importance of conserving the O and N functional groups of purines and pyrimidines at the same positions, 6 and 4 respectively, of the six-membered ring. These strongest preferences can be explained by simple considerations of partial charge interactions between amino acid side chains and base functional groups in the major groove. The basic side chains of arginine and lysine show marked preferences for guanine, which has two negative charges accessible in the major groove at N7 and O6. These significant preferences might be further favored because these long side chains have a relatively high flexibility that might be sufficient to achieve more favorable interactions. The acidic side chain of aspartic acid shows a preference for cytosine, the only base with a net positive charge (at the proton on N4) accessible in the major groove. The asparagine has a side chain with no net charge, but is polar, and it shows a preference for adenine, which also shows no net charge in the major groove (at N6 and N7).

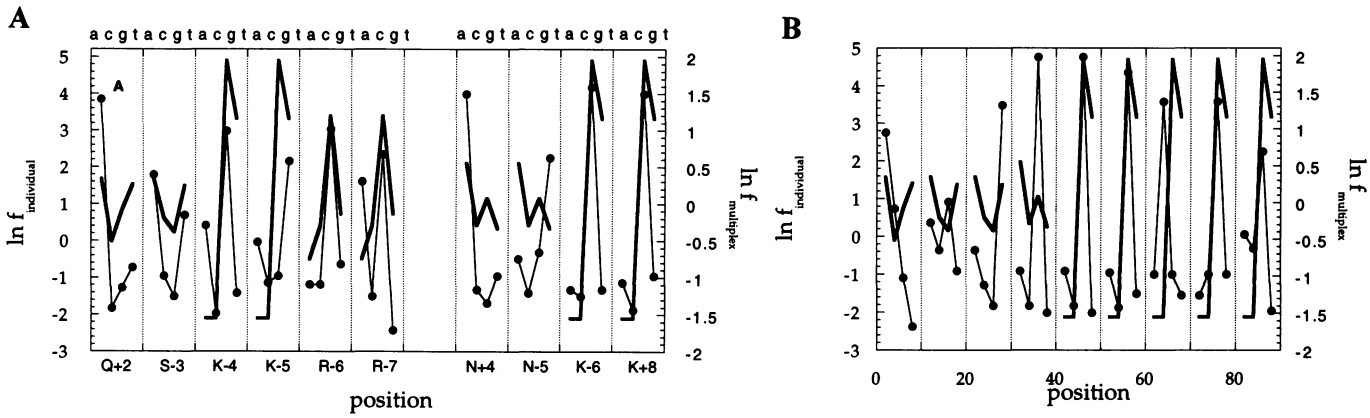

Figure 2. Comparison of relative interaction energies (given as logarithms of frequencies) between different sets of data. Relative interaction energies determined from combinatorial multiplex binding studies (connected by thick lines, without points, indexed on the right ordinate) with individual normalized relative energies (thin lines, with points, indexed on the left ordinate) determined from the binding studies of Takeda *et al.* (1), Sarai and Takeda (2) and Ogata *et al.* (3). (**A**) Left. For substituted consensus portion of OR1 operator DNA bound to Cro protein. Right. For substituted portion of DNA bound to R2R3 region of c-Myb protein. (**B**) For substituted consensus portion OR1 operator DNA bound to λ repressor. Amino acids at fixed positions in the proteins are indexed along the abscissa as in refs 1–3.

Other relative interaction energies

There is some evidence that individual interactions involving a single base-amino acid pair can be treated independently of other interactions. Desjarlais and Berg (7) found that experimentally determined changes in ΔG for various triplet sequences could be predicted from the combinatorial multiplex data by assuming independence of the three base positions in the binding domain. Similarly, free energies of binding for various mutants of OR1 to Cro and λ repressors were predicted by simply adding individual changes in ΔG determined by single substitutions, suggesting that the interactions between a single base and an amino acid can be considered, to first approximation, to be independent (1,2,7). This independence of interaction is utilized here and makes possible calculations of base–amino acid energies from binding studies of substituted DNA, based on the native interactions characterized by X-ray and NMR, for Cro and λ repressors and the R2R3 region of c-Myb protein.

An individual relative interaction energy of a j-type base with a K-type amino acid of fixed position can be expressed in terms of the change in ΔG as

$$\ln f_{Kj} = -(\Delta G_{Kj} - \Delta G_{Ko})/RT$$

where *o* refers to the native base and RT is given as 0.546 kcal/mol (1–3). Each set of four bases is then normalized as in equation **2**. In Figure 2 are shown the comparisons between the relative energies derived from the combinatorial multiplex binding data and the relative interaction energies for OR1 DNA binding to Cro and λ repressors and DNA binding to the R2R3 c-Myb binding domain. The combinatorial binding energies show a significant correlation to these individual energies, although in many cases the magnitudes of the individual interaction energies are larger than those from the combinatorial multiplex data set. Notably, the strongest interactions among the four bases are the same for both the individual energies and the combinatorial multiplex energies, and the overall trends are usually similar for both sets. Linear regression plots for each set indicate slopes near 1 and intercepts near 0, with correlation coefficients of 0.46 for Cro, 0.62 for c-Myb and 0.32 for λ repressor. These results indicate that the probability of any of the three data sets being random (14) is <0.1.

We then calculate the relative frequencies for each of the four bases in the 19 sets of individual contacts shown in Figure 2. These frequencies are then summed for each amino acid type and base type and normalized as in equation **2**. In Figure 3 we compare these aggregate relative energies with those determined from combinatorial multiplex binding for the three sets of bases interacting with arginine, lysine and asparagine. Except for aspartic acid interactions, which are not available, these include the four other strongest interactions that were selected above for further inspection. Perfect agreement would give a linear regression with a slope of 1 and a zero intercept; here we report values of 0.74 and 0.20×10^{-3}, respectively. The correlation coefficient of 0.51 for these 12 points indicates the probability of being random (14) to be <0.1.

A corresponding set of relative interaction energies is derived with equations **1** and **2** from the frequencies of occurrences (1,2,15) for each of the four bases in the six operators of bacteriophage λ, for Cro and λ repressors. The half-operator sequence frequencies combined in ref. 15 have been used. Specification of base–amino-acid contacts relies on the two structures of the consensus half-operators. In Figure 4 we compare these relative interaction energies with the aggregate potentials derived above for Cro, λ repressor and c-Myb. Linear regression indicates a slope of 1.2 and an intercept of 0.33×10^{-3}, with a correlation coefficient of 0.77. There is a clear and strong correlation between the two sets of potentials, with similar strong interactions. These results indicate a strong consistency between the experimental binding constants and the sequence occurrences. In addition, the sequence frequencies appear to reflect directly the DNA–protein interaction energetics.

DISCUSSION

Comparison of sets of relative energies

The extent of correlation between the relative interaction energies calculated from combinatorial multiplex binding of zinc fingers and those derived from the more diverse set of interactions is noteworthy.

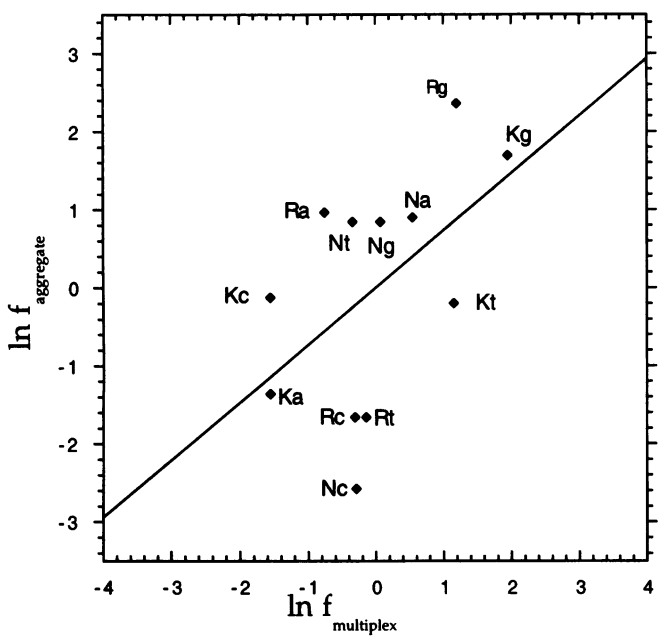

Figure 3. Correlations between two independent data sets derived for different structures and different types of binding data for the observed strongly binding amino acids. Relative interaction energy values are labelled with upper case amino acid and a lower case base designations. Values from multiplex binding studies are plotted (abscissa) against aggregate relative interaction energies (ordinate) determined by averaging over occurrences in specific DNA binding to Cro, c-Myb and λ repressor. The linear regression line is $y = 0.74 \times + 0.20 \times 10^{-3}$; the correlation coefficient is 0.51; $P < 0.10$.

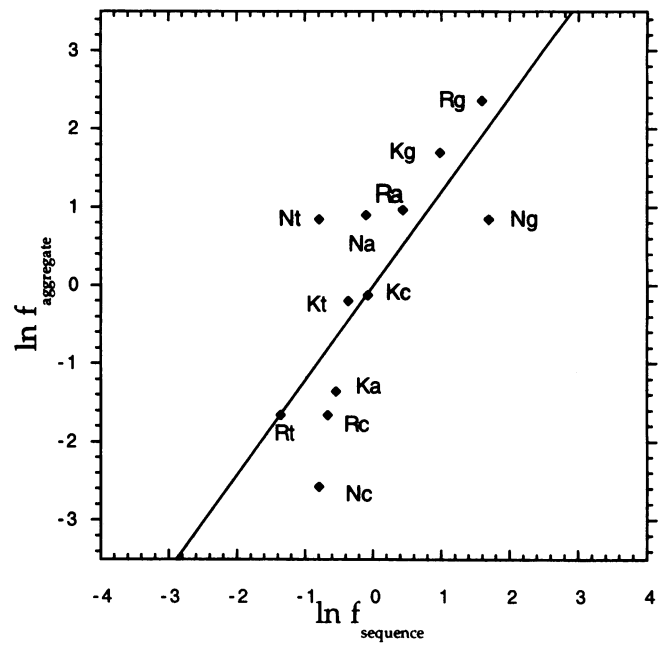

Figure 4. Correlations of interaction energies for most favorably interacting pairs of bases and amino acids derived from sequences with the aggregate binding energy set averaged from values in Figure 2. Relative interaction energies derived from base sequence frequencies for operators of bacteriophage λ (abscissa) are plotted against aggregate relative interaction energies (ordinate). The linear regression line is $y = 1.2 \times +0.33 \times 10^{-3}$; the correlation coefficient is 0.77; $P < 0.50 \times 10^{-2}$.

This correlation supports several important conclusions. First, simple considerations of electric charge can account for the most favorable interaction cases, indicating that hydrophobic interactions are less important for this binding situation. This is not so surprising since there are relatively few non-polar atoms presented in the nucleic acid double helix grooves. Secondly, it appears that physically relevant interaction energies for base–amino acid contacts can be empirically derived directly from combinatorial binding studies with only minimal structure information.

Favorable interactions

The combinatorial multiplex potentials shown in Figure 1 indicate the strongest base preferences to be between arginine–guanine, lysine–guanine and lysine–thymine, and a similar pattern is noted in the aggregate potentials in Figure 3. The most significant base preferences for the residues of the combinatorial multiplex binding can be justified by a simple consideration of partial charge in the major groove of the DNA. It is interesting that electrostatic considerations appear to be critical. However, the role of flexibility of the side chains may also be an additional contributing factor in determining interaction strengths. Finally, it appears that useful potentials can be derived from DNA sequence data, if the sequence positions of strong interactions can be identified.

Combinatorial interaction energies and their application to designing sequence search strategies

One advantage of deriving effective interaction potentials from experimental combinatorial methods is that they can provide large statistical samples that should serve to smooth over less important interactions, as well as individual errors in the data. These potential energies directly characterize the most important interactions of bases and amino acids, reflecting in an empirical way all categories of interactions. The present approach counts every interaction, including multiple interactions of a base or residue. This is consistent with the independence of interactions reported for Cro, λ repressor and zinc finger complexes (1,2,7). Also, such potentials could be derived and updated in the course of a combinatorial binding study, and be used in an iterative fashion to direct the actual experimental substitutions to be biased toward those most likely to stabilize binding.

Development of efficient combinatorial sequence methods to identify functionally important amino acid and nucleotide sequences for binding to specific targets is proving to be a highly significant new area of biotechnology. Such methods usually involve exhaustive searches of large libraries of oligomeric peptides and nucleotides, selecting for molecules with desired binding properties or activities, followed repetitively by sequence refinement. The eventual goal of the approach introduced here is to provide a systematic framework for improving the efficiency of searches over large sequence libraries. This ought ultimately to permit expansion of the searches to include larger target regions, and hence yield more effective binding agents. Current exhaustive sequence searches, although impressive, are still rather limited in the possible sizes of variable ligands.

We suggest that further details of DNA base–amino acid potential energies could be developed, as well as energies for RNA–protein interactions. These potentials should ultimately

improve the efficiencies of experimental combinatorial searches, increasing the feasible sizes and complexities of receptor and ligand molecules. Furthermore, such simplified interaction potentials should prove to be useful tools for molecular design. Applications using similar contact potentials for protein-peptide complexes have proven useful (16,17).

ACKNOWLEDGEMENTS

We thank Drs Victor Zhurkin, Tom Schneider and Jacob V. Maizel, NCI and Dr Akinori Sarai, Tsukuba Life Science Center, for helpful discussions.

REFERENCES

1 Takeda, Y., Sarai, A. and Rivera, V. M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 439–443.
2 Sarai, A. and Takeda, Y. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6513–6517.
3 Ogata, K., Kanei-Ishii, C., Sasaki, M., Hatanaka, H., Nagadoi, A., Enari, M., Nishimura, Y., Ishii, S. and Sarai, A. (1995) submitted.
4 Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y. and Matthews, B. W. (1982) *Nature*, **298**, 718–723.
5 Beamer, L. J. and Pabo, C. O. (1992) *J. Mol. Biol.* **227**, 177–196.
6 Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. and Nishimura, Y. (1994) *Cell* **79**, 639–648.
7 Desjarlais, J. R. and Berg, J. M. (1994) *Proc. Natl. Acad. Sci.* **USA 91**, 11099–11103.
8 Pavletich, N. P. and Pabo, C. O. (1991) *Science* **252**, 809–817.
9 Miyazawa, S. and Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
10 Miyazawa, S. and Jernigan, R. L. (1994) *Protein Eng.* **7**, 1209–1220.
11 Miyazawa, S. and Jernigan, R. L. (1995) submitted.
12 Choo, Y. and Klug, A. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 11163–11167; 11168–11172.
13 Seeman, N. C., Rosenberg, J. M. and Rich, A. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 804–808.
14 Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York.
15 Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188**, 415–431.
16 Young, L., Jernigan, R. L. and Covell, D. G. (1994) *Protein Sci.* **3**, 717–729.
17 Wallqvist, A., Jernigan, R.L. and Covell, D.G. (1995) *Protein Sci.*, **4**, 1881–1903.