

# Structure-Function Relationship in DNA sequence Recognition by Transcription Factors

Akinori Sarai<sup>1</sup>, Samuel Selvaraj<sup>2</sup>, Michael M. Gromiha<sup>3</sup> and Hidetoshi Kono<sup>4</sup>

<sup>1</sup>Dept. Biochemical Engineering and Science, Kyushu Institute of Technology, Iizuka, Japan; sarai@bse.kyutech.ac.jp

<sup>2</sup>Dept. Physics, Bharathidasan University, Tamilnadu, India; selva01@bdu.ernet.in

<sup>3</sup>Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo Japan; michael-gromiha@aist.go.jp

<sup>4</sup>Neutron Science Research Center and Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, Kyoto Japan; kono@apr.jaeri.go.jp

## Abstract

Transcription factors play essential role in the gene regulation in higher organisms, binding to multiple target sequences and regulating multiple genes in a complex manner. In order to understand the molecular mechanism of target recognition, and to predict target genes for transcription factors at the genome level, it is important to analyze the relationship between the structure and function (specificity) of transcription factors. We have used a knowledge-based approach, utilizing rapidly increasing structural data of protein-DNA complexes, to derive empirical potential functions for the specific interactions between bases and amino acids as well as for DNA conformation, from the statistical analyses on the structural data. Then these statistical potentials are used to quantify the specificity of protein-DNA recognition. The quantification of specificity has enabled us to establish the structure-function analysis of transcription factors, such as the effects of binding cooperativity on target recognition. The method is also applied to real genome sequences, predicting potential target sites. We are also using computer simulations of protein-DNA interactions in order to complement the empirical method. Combining the two approaches together, we can better understand the mechanism of protein-DNA recognition and improve the target prediction of transcription factors.

*Keywords:* protein-DNA recognition; structure-function relation; transcription factors; target prediction

## 1 Introduction

Regulation of gene expression in higher organisms is achieved by a specific recognition of target DNA sequences by DNA-binding proteins. Due to the progress of X-ray crystallography and NMR spectroscopy techniques, structural data on the protein-DNA

complexes have been rapidly increasing. However, the mechanism of DNA sequence recognition by proteins has been poorly understood, and thus the accurate prediction of their targets at the genome level is not yet possible. This situation implies that the structural information has not been fully utilized. Understanding the molecular mechanism and its application to genome-wide prediction are essential for the analysis of gene regulation network. Here, we describe two kinds of approaches for studying protein-DNA recognition. One is a knowledge-based approach, by which we extract functional information from structural data of protein-DNA complexes. We made a statistical analysis of structural database of protein-DNA complex, and derived empirical potential functions for the specific interactions between bases and amino acids. Then, we used a sequence-structure threading to examine the relationship between structure and specificity in protein-DNA recognition, and to predict target sites of transcription factors in real genome sequences. We also evaluated the fitness of DNA sequence against DNA structure to examine the role of indirect readout mechanism. We show how the structural features are related to the specificity, and discuss relative roles of direct and indirect readout mechanisms in the recognition. We also discuss the strategy to predict the target sequences of transcription factors at the genome level, and its application to yeast genome. In another approach, we analyze protein-DNA recognition by computer simulations. We describe several different methods of computer simulations.

## 2 Methods

We extracted interacting pairs of bases and amino acids from a set of non-redundant protein-DNA complex structures (Kono and Sarai, 1999; Selvaraj et al., 2002). In order to derive the statistical potential of interactions between bases and amino acids, we analyzed the amino acid distributions around each base. We defined a coordinate system by taking an origin N9 atom for A and G and N1 atom for T and C. We considered the amino acids within a given box, and the box was divided into grids. Because the sample numbers are not yet very large, we first consider the information about C<sub>α</sub> atoms. Then we transformed the distributions of C<sub>α</sub> atom of amino

acid into statistical potentials defined by the following equations (Sipl, 1990)

$$\begin{aligned} \Delta E^{ab}(s) &= RT \ln \frac{f^{ab}(s)}{f(s)} \\ f^{ab}(s) &= \frac{1}{1 + m_{ab}w} f(s) + \frac{m_{ab}w}{1 + m_{ab}w} g^{ab}(s) \end{aligned}$$

where  $m_{ab}$  is the number of pairs, amino acid  $a$  and base  $b$  observed,  $w$  is the weight given to each observation,  $f(s)$  is the relative frequency of occurrence of any amino acids at grid point  $s$ , and  $g^{ab}(s)$  is the equivalent relative frequency of occurrence of amino acid  $a$  against base  $b$ .  $R$  and  $T$  are gas constant and absolute temperature, respectively. Here, we used a box of  $|x| = |y| = 13.5 \text{ \AA}$  and  $|z| = 6 \text{ \AA}$  and a grid interval of  $3 \text{ \AA}$ , which was determined by examining various intervals.

By threading a set of random DNA sequences onto the template structure, we calculated the Z-score of the specific sequences against the random sequences, which represents the specificity of the complex. Assuming the additivity of potential energies, the sum of the potential energies ( $E_{PD} = \sum_{ab,s} \Delta E^{ab}(s)$ ) for a given DNA sequence in a complexed form was defined as the energy for the sequences. The energy for a particular sequence, in a crystal structure for example, was normalized to measure specificity by the Z-score against random sequences. The Z-score was defined as  $(X - m)/\sigma$ , where  $X$  is the energy of a particular sequence,  $m$  is the mean energy of 50,000 random DNA sequences, and  $\sigma$  is the standard deviation.

We have also derived statistical potential functions for conformational energy of DNA from the protein-DNA complex structural data to evaluate the fitness of sequences to a particular conformation of DNA. To estimate the sequence-dependent DNA conformational energy, we mostly followed the approach described by Olson et al. (1998). The conformation energies were approximated using a harmonic function,  $E_{DNA} = 1/2 \sum_{ij} f_{ij} (\Delta_{ij} - \Delta_{ij}^0)^2$ , in which  $\Delta_{ij}$  represents the base-step parameters, and  $f_{ij}$  are the elastic force constants impeding deformation of the given base step and  $\Delta_{ij} = \Delta_{ij} - \Delta_{ij}^0$ , in which  $\Delta_{ij}^0$  is the average base-step parameter. The base-step parameters used were shift, slide, rise, tilt, roll, and twist. The unknown parameters  $f_{ij}$  and  $\Delta_{ij}^0$  were determined by statistical analysis of the same non-redundant protein-DNA complexes. Setting up a covariance matrix from observed distributions of  $\Delta_{ij}$  thus refers to an effective inverse harmonic force-constant matrix. Inversion of this matrix transformed it to a force-constant matrix in the original coordinate basis. We removed all parameters of a base step for which one parameter exceeded three standard deviations, in an iterative manner. Then the final force field was calculated. The conformational energy of DNA in a given complex structure was calculated as the sum of all the base steps. We assigned the energy corresponding to the threshold value when any parameter exceeded three standard deviations. Then, these potentials were used to quantify the specificity of indirect readout

mechanism of protein-DNA recognition, as a Z-score, by using the same threading procedure.

We carried out jack-knife and bootstrap tests to assess the statistical confidence in the Z-score calculations. To remove the effect of self contributions, we always removed the self from the original dataset of complex structures when calculating its Z-score. We then examined the Z-score further by removing one additional randomly selected structure from the dataset and repeated this procedure. We found that the Z-scores were stable against these treatments, indicating that our dataset of protein-DNA complexes provides adequate information for a generally valid structure-based potential. To calculate the bootstrap standard errors, we prepared a set of 61 randomly selected complex structures in which the self was removed but duplications of the same structure were allowed. We created 200 such replications and calculated the standard errors.

We combined the direct and indirect energies to calculate the total energy. Because the derivations of these empirical energies are based on different statistics, we cannot simply make a summation. We need to introduce a weighting factor:  $E_{tot} = cE_{PD} + (1-c)E_{DNA}$ , where  $E_{PD}$  and  $E_{DNA}$  are the energies of the direct and indirect readout, respectively, and  $c$  is a weighting coefficient ranging between 0 and 1. This coefficient is determined by maximizing the total Z-score – i.e., the Z-score is calculated from random sequences, and a value of  $c$  is sought that gives the highest total Z-score. Then, the threading procedure was applied to the real genome sequences in order to find potential target sites, by using the total energy.

We carry out several different types of computer simulations depending on the resolution of the system under investigation: Monte Carlo simulation of base amino acid interactions; molecular dynamics and free energy calculations of protein-DNA complex; empirical calculations of interaction energy of protein-DNA complex; and docking simulation of protein-DNA binding and sliding.

### 3 Structure-Specificity Relationship in Protein-DNA Recognition Revealed by Statistical Potentials

#### 3.1 Cognate/noncognate, Symmetry/asymmetry and Cooperative Binding

It is interesting to compare two structures, cognate and non-cognate complex structures in order to understand what is important for specific binding and what is different between them. There are several such examples in PDB (Protein Data Bank). We examined NF- $\kappa$ B, glucocorticoid receptor DNA binding domain, EcoRV endonuclease and BamHI endonuclease, for which both the cognate and non-cognate complex structures are available in PDB. The statistical method could distinguish the two structures as differences in the Z-scores as well as statistical potentials (Kono and Sarai, 1999; Selvaraj et al., 2002). Thus, the subtle differences

in specificity of these structures could be detected by the analysis of energies. Fig. 1 shows an example of NF- $\kappa$ B.

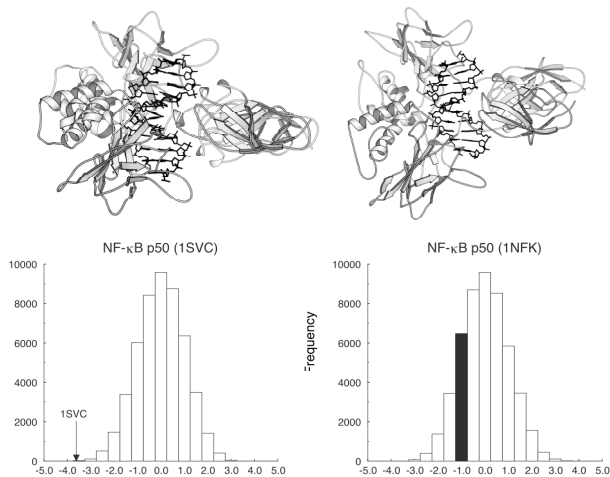


Fig. 1. Structures of NF- $\kappa$ B p50 complexed with 11-bp consensus sequence (PDB: 1SVC) (left) and with 10-bp DNA (PDB: 1NFK) (right). The binding histograms show Z-scores of respective complexes (arrow and filled bar).

Proteins often bind to DNA as homodimers, which leads to subtle structural differences between the two subunits. Thus, we examined in detail the structural effects of asymmetric binding on specificity. Marked differences in the specificity of DNA binding were observed for the two subunits of  $\lambda$  repressor, the glucocorticoid receptor, and for transcription factors containing a  $Zn_2Cys_6$  binuclear cluster domain, which are known to bind asymmetrically to DNA (Selvaraj et al., 2002). Table 1 shows some examples of asymmetric Z-scores for homodimeric protein-DNA complexes.

Protein-DNA complex	Z-scores		
	whole	chain1	chain2
Pyrimidine pathway regulator	-2.9	-3.1	-1.1
$\lambda$ repressor	-2.6	-2.9	-0.1
Glucocorticoid Receptor	-1.1	-1.8	0.6

Table 1. Z-scores for homodimeric protein-DNA complexes. Z-scores were calculated for the whole complex and two monomers separately.

We also applied this method to examine the relationship between structure and specificity in cooperative protein-DNA binding. The effect of cooperative binding was examined by comparing the monomer and heterodimer complexes of MATA1/ $\lambda$ 2 (Kono and Sarai, 1999), MCM1/MAT $\lambda$ 2 and NFAT/AP-1 transcription factors. We found that the heterodimer binding enhances the specificity in a non-additive manner. This result indicates that the conformational changes introduced by the heterodimer binding play an important role in enhancing the specificity. The structures of cognate and noncognate complexes of EcoRV show marked differences in the conformations of

both the enzyme and DNA. The example of EcoRV enabled us to show the cooperative effect of sequence and structure on specificity  $\square$  i.e., conformational differences between the cognate and noncognate complexes were sensitive to the cognate but not the noncognate DNA sequence, and the cognate structure had a greater ability to discriminate DNA sequences than the noncognate structure (Selvaraj et al., 2002).

### 3.2 Direct/Indirect Readout

The above method is based on the direct readout mechanism, in which protein recognizes DNA sequence through the direct contact between amino acids and base pairs. On the other hand, substitutions of those base pairs not in contact with amino acids often affect binding affinity, indicating that protein may recognize DNA sequence through particular structure or property of DNA. This indirect readout mechanism may contribute to the specificity of protein-DNA recognition significantly. We have derived statistical potential functions for conformational energy of DNA to quantify the specificity of indirect readout mechanism of protein-DNA recognition (Sarai et al., 2001, Gromiha et al., 2003). Once the potentials are derived, the conformational energy of DNA sequence can be estimated for given structure, and the threading procedure can be used to evaluate the fitness of sequence to structure of DNA. We can calculate the Z-score for the indirect recognition in the same way as for the direct recognition. By comparing both the Z-scores we can assess the relative contributions of direct and indirect readout mechanisms. We have analyzed various protein-DNA complexes systematically, and found that both the direct and indirect mechanisms make significant contribution to the specificity (Gromiha et al., 2003). The relative contributions depend on the types of DNA-binding proteins.

Because both the potentials are independent quantities, they can be summed up to calculate the total energy and used to find target sites, although a weighting factor needs to be determined as the two potentials were

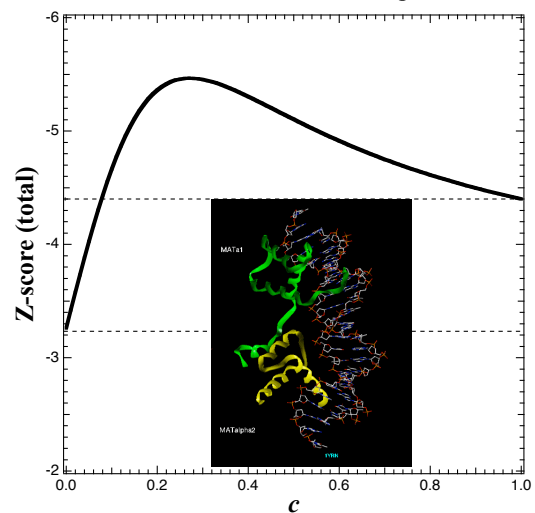


Fig. 2. Total Z-score with respect to weight factor,  $c$ , for MATA1/ $\lambda$ 2-DNA complex. Total energy is given by  $E_{tot} = cE_{protein-DNA} + (1-c)E_{DNA}$ .

derived from different statistics. Figure 2 shows the Z-score calculated for the total energy with varying weighting coefficient. One can see enhanced Z-score compared with individual Z-scores for direct or indirect readout ( $c=1$  or 0, respectively). This result indicates that the energies of the direct and indirect readouts contain independent information that in combination enhances the specificity of the recognition. We use the weight coefficient that gives the maximum total Z-score. We use the combined energy for threading against DNA sequences to make target predictions for transcription factors.

### 3.3 Target Prediction in Yeast Genome

The threading procedure was used to find target sites of transcription factors in real genome sequences. As an example of such applications, we could identify the experimentally-verified binding sites of the transcription factor MATa1/□2 in the promoter of *HO* gene successfully (Kono and Sarai, 1999). We have also attempted to identify target sites and genes of MCM1/MAT□2 in the whole yeast genome. The target genes of this transcription factor have been identified in yeast genome experimentally (Zhong et al, 1999). The predicted target genes were ranked by the Z-score and compared with experimental data. The target genes identified positively by experiment were ranked high in the list, and the experimentally negative genes were ranked low. Separation between the positive and negative genes was not perfect but they were segregated by a certain threshold Z-score value. The total Z-score gave better separation than that of direct contribution alone. We are applying the method for the targets prediction of other transcription factors.

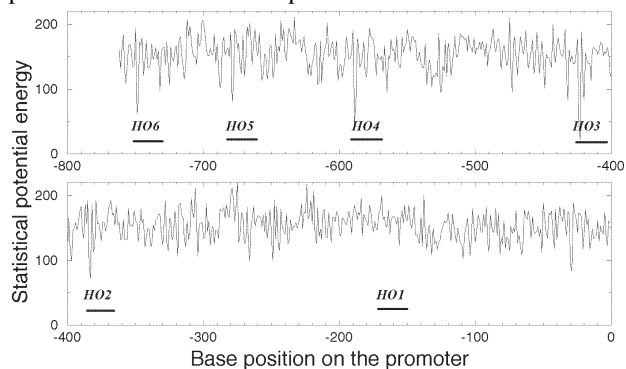


Fig. 3. Prediction of the binding sites by a MATa1/□2 complex in the upstream promoter region of *S. cerevisiae* site-specific endonuclease (*HO*) gene. The sharp peaks are the predicted binding sites. HO3-HO6 were shown to be strong binding sites.

## 4 Computer Simulations of Protein-DNA Interactions

The statistical method has shown that the distribution of  $C_{\square}$  position of amino acids around base pairs provides important information about the specificity in the DNA sequence recognition by proteins. However, the accuracy of this prediction method is limited by the number of available structural data. Thus, it is desirable to

complement the method by some other means. Computer simulation is one such method. In real DNA-protein interactions, however, there are many factors contributing to the recognition process. Thus, it is necessary to examine different levels of the system by different methods. Below we describe results from different kinds of computer simulations.

### 4.1 Free-Energy Map for Base-Amino Acid Interactions

In the case of base-amino acid interactions, we need to reproduce the distribution of  $C_{\square}$  position of amino acids around base pairs observed in the experimental DNA-protein complex structures. Thus, it would be reasonable to consider at first the interactions between a base pair and an amino acid. In reality, the  $C_{\square}$  position is fixed by main chain of protein and the possible range of  $C_{\square}$  direction may be restricted. However, such biases, which are context dependent, are difficult to evaluate *a priori*. Therefore, at first we will consider intrinsic interactions between a base pairs and an amino acid. We generated many  $C_{\square}$  positions and side chain conformations by systematic sampling (Pichierri et al., 1999; Yoshida et al., 2002) or Monte Carlo sampling methods (Sayano et al., 2000), and calculated free energy map of  $C_{\square}$  around a given base pair. By calculating the free energies for different  $C_{\square}$  positions and subtracting a reference free energy at a large separation, we can obtain a contour map of interaction free energy, which shows preferable positions of  $C_{\square}$  of amino acid around a base pair. This can be directly compared with the distribution of  $C_{\square}$  position of amino acids around base pairs derived from DNA-protein complex database. According to the free-energy contour maps for the interactions of Asn with A-T and with G-C, the preferable position of  $C_{\square}$  is localized in a narrow region around A in the case of A-T (Fig. 3). In this region, Asn and A form specific double hydrogen bonds,  $C=O \cdots HN6$  and  $NH \cdots N7$ , which are found frequently in the Asn-A pair in the experimental structures of DNA-protein complexes. Also, the

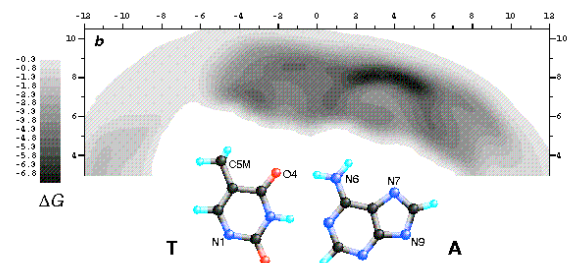


Fig. 4. Free-energy maps of the interaction between Asn and A-T. Darker region corresponds to low free energy, where  $C_{\square}$  position of Asn is stable.

distribution of  $C_{\square}$  is in agreement with the statistical potential obtained by the database analysis (Kono and Sarai, 1999). On the other hand, Asn tends to be more broadly distributed around G-C. The lowest  $\Delta G$  values are located in the middle of G-C, and the following lower  $\Delta G$ 's extend towards C5 atom of C, where C does

not have a methyl group. This comparison indicates that the interaction of Asn is more specific toward A-T than G-C. This example illustrates how we can quantify the specificity in the base-amino acid interactions and complement the statistical method.

## 4.2 Semi-Empirical Estimation of Free-Energy Changes due to Base Mutations

The specificity of protein-DNA recognition is usually assessed by the effect of mutations. If mutations cause large effect on the binding affinity, it would indicate that specific interactions are involved at the mutation site, and the extent of specificity dictates the effect of mutation. Thus, it is important to examine the effect of mutation and component of interaction energy causing the effect, as well as to develop methods to predict the mutation effect. We have performed a computer analysis in which a phage DNA-binding protein,  $\lambda$  repressor, was used to examine the changes in binding free-energy ( $\Delta G$ ) and its energy components caused by single base mutations (Oobatake et al., 2003). We then determined which of the calculated energy components best correlated with the experimental data (Sarai and Takeda, 1989). The experimental  $\Delta G$  values were well reproduced by the calculations. Component-analysis revealed that the electrostatic and hydrogen bond energies were most strongly correlated with the experimental data. Among the 51 single base-substitution mutants examined, positive  $\Delta G$  values, corresponding to weakened binding, were caused by the loss of favorable electrostatic interactions and hydrogen bonds, the introduction of steric collisions and electrostatic repulsion, the loss of favorable interactions with a thymine methyl group, and the increase of unfavorable hydration energy from isolated DNA. These results suggest that electrostatic interactions and hydrogen bond make major contributions to the specificity of  $\lambda$  repressor-DNA recognition. However, van der Waals and hydrophobic interactions also play important role in particular DNA sequence locations.

## 4.3 Free-Energy Calculations of Protein-DNA Complexes

For a large system like  $\lambda$  repressor DNA complex, it is not an easy task to use all-atom simulations to calculate interaction free energy changes caused by mutations accurately. However, calculation of free energy changes due to small structural perturbation in the complex is feasible. We calculated a binding free energy change between DNA and  $\lambda$  repressor due to a base substitution from thymine (T) to uracil (U) by the free energy perturbation method based on molecular dynamics simulations for the complex in water with all degrees of freedom and long-range Coulomb interactions (Saito and Sarai, 2003). The binding free energy change calculated was in good agreement with an experimental value (calculated and experimental values of  $\Delta G$  are, respectively, 1.5 kcal/mol and 1.8 kcal/mol, Sarai and Takeda, 1989). By using component analysis, we could clarify the reason why the small difference between T and U ( $\text{CH}_3$  in T is replaced by H in U) caused the

significant binding free energy change. The free energy change is due to the gain of hydration free energy in the dissociated state and the loss of favorable van der Waals interactions in the associated state.

## 4.4 Docking Simulations of Protein-DNA Complexes

During the recognition process, proteins associate and dissociate with DNA, and may slide along DNA before finding their target sites. Thus, protein-DNA recognition is a dynamical process. In order to reveal reaction pathways of proteins along DNA, we are conducting docking simulations of protein-DNA complexes by using Monte Carlo method. At first, using rigid-body approximations, we examined free-energy profile along the DNA surface for homeodomains and restriction enzymes. Preliminary results indicate that the homeodomain may track along the major groove of DNA while it slides on DNA. We will incorporate the flexibilities of protein and DNA molecules such as hinge and bending motions, and examine the role of these effects on the protein-DNA recognition process.

## 5 Discussion

We have described two kinds of methods for studying the relationship between structure and function (specificity) in protein-DNA recognition. One is the knowledge-based approach, by which we extract biological information from structural data on protein-DNA complexes. The increase in the structural data, which will be accelerated by structural genomics project, will make this structure-based method promising for revealing the structure-function relationship in protein-DNA recognition and for predicting targets of transcription factors. This method can also be applied to proteins of unknown structure having substantial sequence similarity to known proteins: the structure can be modeled based on the similarity and its binding sites can be predicted. This approach is, however, limited by the number of available data. Thus, we try to complement it with a deductive approach, by which we try to reproduce the recognition process by computer simulation. Although this approach requires a large amount of computational time, the advancement of technology makes the computer simulation of large systems feasible. The information obtained by this approach would help the knowledge-based approach in improving the accuracy of statistical potentials and target prediction. Thus, a combination of the two methods will become a powerful tool for the study of protein-DNA recognition and its application to target prediction at the genome level.

## Acknowledgements

This work was partly supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports and Technology awarded to A.S. (15014234) and H.K. (15014232).

## 6 References

- Gromiha, M., Siebers, J.G., Selvaraj, S., Kono, H. and Sarai, A. (2003): Direct or Indirect Readout in Protein-DNA Recognition?" *J. Mol. Biol.* submitted.
- Kono, H. and Sarai, A. (1999): Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* 35, 114-131.
- Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. and Zhurkin, V.B. (1998): DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA* 95, 11163-11168.
- Oobatake, M., Kono, H., Wang, Y-F. and Sarai, A. (2003): Anatomy of specific interactions between lambda repressor and operator DNA by energy-component analysis. *Proteins*, 53, 33-43.
- Pichierri, F., Aida, M., Gromiha, M. and Sarai, A. (1999): Free energy maps of base-amino acid interaction for protein-DNA recognition. *J. Am. Chem. Soc.* 121, 6152-6157.
- Saito, M. and Sarai, A. (2003): Free-energy-perturbation calculations of repressor-DNA complex" *Proteins* 52, 129-136.
- Sarai, A. and Takeda, Y. (1989): Lambda Repressor Recognizes the Approximately 2-fold Symmetric Half-Operator Sequences Asymmetrically. *Proc. Natl. Acad. Sci. USA*. 86, 6513-6517.
- Sarai, A., Selvaraj, S., Gromiha, M.M., Siebers, J.G., Prabakaran, P. & Kono, H. (2001): Target prediction of transcription factors: refinement of structure-based method. *Genome Informatics* 12, 384-385.
- Sayano, K., Kono, H., Gromiha M. and Sarai, A. (2000): Multicanonical Monte Carlo Calculation of Free-Energy Map for Base-Amino Acid Interaction. *J. Compt. Chem.* 21, 954-962.
- Selvaraj, S., Kono, H. and Sarai, A. (2002): Specificity of Protein-DNA Recognition Revealed by Structure-Based Potentials: Symmetric/Asymmetric and Cognate/Noncognate Binding. *J. Mol. Biol.* 322, 907-915.
- Sippl, M. (1990): Calculation of conformational ensembles for potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213: 859-883,.
- Yoshida, T., Nishimura, T., Aida, M., Pichierri, F., Gromiha, M.M. and Sarai, A. (2002): Evaluation of free energy landscape for base-amino acid interactions using ab initio force field and extensive sampling. *Biopolymers* 61, 84-95.
- Zhong, H., McCord, R. and Vershon, A.K. (1999): Identification of target sites of the  $\lambda$ 2-Mcm-1 repressor complex in the yeast genome. *Genome Res.* 9:1040-1047,.