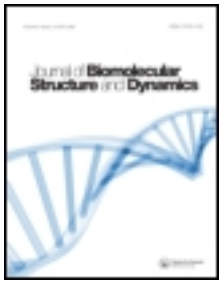

Journal of Biomolecular Structure and Dynamics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

Recognition rules for binding of Zn-Cys2His2 transcription factors to operator DNA

R.V. Polozov[a], V.S. Sivozhelezov[bc], Yu.N. Chirgadze[d] & V.V. Ivanov[e]

[a] Institute of Theoretical Experimental Biophysics, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia

[b] Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia

[c] Nanoworld Institute, Fondazione ELBA Nicolini, Pradalunga, Bergamo 24052, Italy

[d] Institute of Protein Research, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia

[e] Joint Institute for Nuclear Research, Dubna 141980, Moscow Region, Russia

Published online: 27 Jan 2014.

To cite this article: R.V. Polozov, V.S. Sivozhelezov, Yu.N. Chirgadze & V.V. Ivanov , Journal of Biomolecular Structure and Dynamics (2014): Recognition rules for binding of Zn-Cys2His2 transcription factors to operator DNA, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2013.879074

To link to this article: http://dx.doi.org/10.1080/07391102.2013.879074

Journal of Biomolecular Structure and Dynamics, 2014
http://dx.doi.org/10.1080/07391102.2013.879074

Recognition rules for binding of Zn-Cys2His2 transcription factors to operator DNA

R.V. Polozov[a], V.S. Sivozhelezov[b,c], Yu.N. Chirgadze[d]* and V.V. Ivanov[e]

[a]*Institute of Theoretical Experimental Biophysics, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia;* [b]*Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia;* [c]*Nanoworld Institute, Fondazione ELBA Nicolini, Pradalunga, Bergamo 24052, Italy;* [d]*Institute of Protein Research, Russian Academy of Sciences, Pushchino 142290, Moscow Region, Russia;* [e]*Joint Institute for Nuclear Research, Dubna 141980, Moscow Region, Russia*

Communicated by Ramaswamy H. Sarma

(*Received 21 October 2013; accepted 23 December 2013*)

The molecules of Zn-finger transcription factors consist of several similar small protein units. We analyzed the crystal structures 46 basic units of 22 complexes of Zn-Cys2His2 family with the fragments of operator DNA. We showed that the recognition of DNA occurs via five protein contacts. The canonical binding positions of the recognizing α-helix were −1, 3, 6, and 7, which make contacts with the tetra-nucleotide sequence ZXYZ of the coding DNA strand; here the canonical binding triplet is underlined. The non-coding DNA strand forms only one contact at α-helix position 2. We have discovered that there is a single highly conservative contact His7α with the phosphate group of nucleotide Z, which precedes each triplet XYZ of the coding DNA chain. This particular contact is invariant for the all Zn-Cys2His2 family with high frequency of occurrence 83%, which we considered as an invariant recognition rule. We have also selected a previously unreported Zn-Cys2His2-Arg subfamily of 21 Zn-finger units bound with DNA triplets, which make two invariant contacts with residues Arg6α and His7α with the coding DNA chain. These contacts show frequency of occurrence 100 and 90%, and are invariant recognition rule. Three other variable protein-DNA contacts are formed mainly with the bases and specify the recognition patterns of individual factor units. The revealed recognition rules are inherent for the Zn-Cys2His2 family and Zn-Cys2His2-Arg subfamily of different taxonomic groups and can distinguish members of these families from any other family of transcription factors.

Keywords: protein-DNA recognition; transcription factor; recognition rules; DNA binding protein family; Zn-finger; invariant contacts; variable contacts

Introduction

A transcription factor is a protein that binds to a specific sequence of operator DNA, thereby controlling the gene transcription. More than 5% of human genes are known to encode the transcription factors that reflect their biological importance in the cellular functioning. Protein-DNA recognition is an initial key stage, which is followed by several steps. At large distances, the interaction is determined by the electrostatic fields of protein and DNA, which provide for the necessary positions and orientation of the protein along the major groove of the operator DNA. At short interatomic distances, a number of specific contacts are formed between atoms of protein and DNA. Such alternation of analog and digital aspects of binding appears to be a distinctive feature of the recognition process (Chirgadze, Zheltukhin, Polozov, Sivozhelezov, & Ivanov, 2009). Recognition occurs essentially in the major groove of the DNA molecule through the contacts between amino acid side chains of the protein and bases and phosphates of the DNA molecule. The interaction in the minor groove of DNA usually includes fewer contacts. At present, many aspects of the protein-DNA binding, such as classification of the packing interfaces, identification of thermodynamic and kinetic parameters of the protein-DNA complex formation, and deciphering of mechanisms of complex formation, are still largely unsolved (Choo & Klug, 1997; Klug, 2010; Rhodes, Schwabe, Chapman, & Fairall, 1996; Surai & Kono, 2005). Particularly, the detailed analysis of atomic contacts in the interface regions of protein-DNA complexes is far from being completed (Rohs et al., 2010).

Currently, different DNA-recognition proteins are divided into 71 structural super families and 207 families according to the SCOP database (Murzin, Brenner, Hubbard, & Chothia, 1995). There are over 2000 solved crystal structures and over 100 solution NMR structures of protein-DNA complexes of various types. In fact, a great variety of data provides a basis for a thorough comparative analysis of spatial complex structures and their interfaces, with the ultimate goal to find the DNA-protein recognition rules for definite kinds of binding

*Corresponding author. Email: chir@vega.protres.ru

factors. All these data are a basis for a detailed per family analysis of transcription factors to deduce the recognition rules within each family of DNA-recognizing proteins.

The discovery of recognition rules for the protein-DNA binding which is specific for each protein factor family is of great importance and continued interest (Berg & Shi, 1996; Choo & Klug, 1997; Freemont, Lane, & Sanderson, 1991; McCammon, 1998; Rhodes et al., 1996; Suzuki, Gerstein, & Yagi, 1994; Surai & Kono, 2005). Previously, there was a heated discussion on the existence of the protein-DNA recognition code decoding the direct relationships between amino acids and nucleotides (Choo & Klug, 1997; Matthews, 1988; Suzuki, Brenner, Gerstein, & Yagi, 1995). In fact, some important protein-DNA binding relationships between arginine and guanine or between asparagine and adenine were identified for different kinds of protein-DNA complexes. However, the general code was inferred only probabilistically (Benos, Lapedes, & Stormo, 2002).

We hypothesize that the interfaces of complexes from different DNA-binding protein families are organized according to certain recognition rules. This assumption is supported by many empirical and theoretical arguments discussed in most of the papers. We have analyzed the interfaces in homeodomain-DNA complexes and discriminated contacts by their significant types and, as a result, deduced the common recognition rules for this protein family (Chirgadze, Sivozhelezov, Polozov, Stepanenko, & Ivanov, 2012). Hence, in homeodomains, there are about 20 protein-DNA binding contacts, which are localized in the major groove. The systematic study of homeodomain-DNA binding interfaces allowed us to discover a few invariant contacts, which include six contacts between the transcription factor and coding and non-coding DNA chains. In particular, three polar residues (Trp2, Asn5, and Arg7) of the recognizing α-helix form contacts with very high frequency of occurrence (95–100%) with both DNA strands. As a result, we deduced the recognition rules for the whole homeodomain family, which include different taxonomic groups. It was shown that these rules could distinguish members of this family from those of any other family of transcription factors. Our study of homeodomain recognition encouraged us to investigate the Zn-finger transcription factor family. Until now, attempts to formulate common rules for Zn-finger factors were described elsewhere (Ponomarenko et al., 1999; Paillard & Lavery, 2004; Reddy, Das, & Jayaram, 2001; Steffen, Murphy, Tolleri, Hatfield, & Lathrop, 2002; Surai & Kono, 2005). These factors are used to design important artificial proteins to recognize binding specific DNA sequences. However, most studies were devoted solely to the well-characterized mouse Zif268 complexes, which contained three Zn-finger units.

Here, we have studied structures of Zn-finger complexes with fragments of operator DNA. The Zn-finger family is very different from homeodomains in terms of their structures and types of protein-DNA binding patterns (Andreini, Banci, Bertini, & Rosato, 2006; Auld, 2001; Christianson, 1991; Gamsjaeger, Liew, Loughlin, Crossley, & Mackay, 2007; Laity, Lee, & Wright, 2001; Suzuki et al., 1994; ; Wolfe, Nekludova, & Pabo, 2000). The basic ββα motif of the unit Zn-finger factors contains long α-helix of 12 residues, termed the DNA-recognizing helix. The Zn-ion is coordinated by two cysteines and two histidines. This family is called Zn-Cys2His2. Usually, the zinc transcription factor proteins contain multiple zinc finger modules, which form a tandem super structure along DNA in the major groove. Each finger forms a conservative ββα motif, and amino acids on the surface of the so-called recognizing α-helix make contacts with the bases and phosphates of DNA. The module contains a Zn atom and is similar to metalloprotein complex. The most specific feature of the structure of Zn-finger factors is that they include from two to nine, and sometimes more, almost identical Zn-finger units. Such a specific spatial design and combinatory principle of recognition implies a variety of Zn-finger protein factors, which have a broader range of functions than the homeodomains (Desjarlais & Berg, 1992; Krishna, Majumdar, & Grishin, 2003; Laity et al., 2001; Wolfe et al., 2000).

The modular principle of Zn-finger transcription factors was discovered 30 years ago by studying the transcription factor TFIIIA of the frog oocytes from *Xenopus laevis,* which works as an activator and regulator of the gene of ribosomal 5S RNA (Miller, McLachlan, & Klug, 1985). This factor consists of nine sequential Zn-finger units specifically bound to a double-chained DNA of about 50 nucleotide pairs. The main features of this Zn-finger transcription factor are shown in a simplified form in Figure 1. The first atomic crystal structure of a complex, containing three zinc fingers from Zif268 (a mouse immediate early protein) and a consensus DNA-binding site, was determined at 2.1 Å resolution about 20 years ago (Pavletich & Pabo, 1991). Later, the structure of this complex at a much higher resolution 1.6 Å was obtained (Elrod-Erickson, Rould, Nekludova, & Pabo, 1996). The details of this remarkable structure can be seen in Figure 2. At present, Zn-finger factors attract increasing attention in relation to the challenging task: the design of DNA-binding proteins to control gene expression (Klug, 2010). Spatial structures of many complicated complex Zn-finger factors, mostly mouse and human ones, have been reported during the last decade.

Zn-finger transcription factors are one of the most common families of eukaryotic proteins that recognize DNA; and they are rarely found in prokaryotes. These

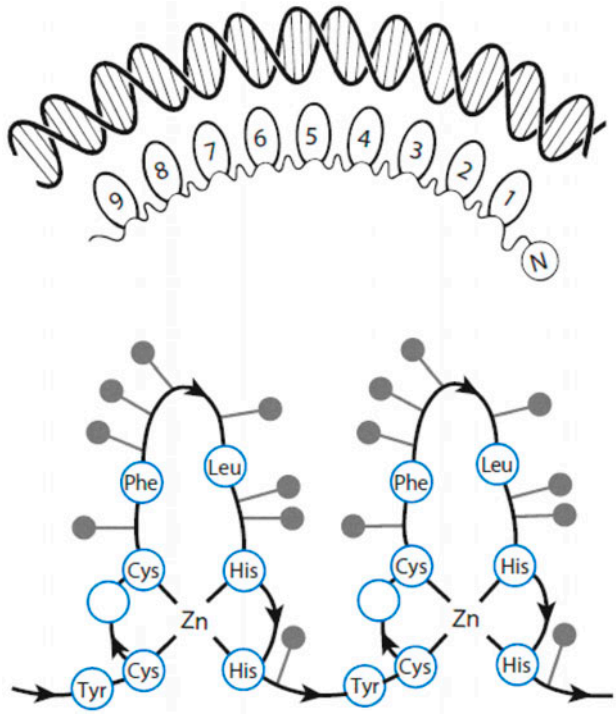

Figure 1. Schematic diagram of repeated Zn-finger protein unit as it was discovered in the study of transcription factor TFIIIA from oocytes *Xenopus laevis* (Miller et al., 1985). The upper part of figure shows double-stranded DNA and modular structure of transcription factors; the numbers are related to protein factor units. Coordination of Zn by amino acids in two adjacent protein units is shown at the lower part of figure.

proteins play an important role in many cellular processes including replication, transcription, translation, repair, metabolism, cell proliferation, and apoptosis. Initially, Rhodes et al. (1996) classified zinc fingers. At present the proteins from this family are assigned to eight different folds, corresponding to three main types of domains: Cys2His2, Treble Clef, and Ribbon (Rohs et al., 2010). These types of the domains are present not only in protein-DNA complexes, but also in protein–RNA and protein–lipid complexes. The most common coordination of zinc ions in these proteins is tetrahedral, but pentahedral or hexahedral coordination also occurs (Patel, Kumar, & Durani, 2007). A replacement of cysteine or histidine in the coordination sphere leads to the loss of protein function (Patel et al., 2007). As a rule, there are no water molecules in the Zn-coordination sphere (Auld, 2001). Interestingly, zinc finger modules have only a limited number of conserved amino acid residues.

Until now, no recognition rules for binding the Zn-finger to DNA have been found. One reason of this is that the amount of suitable data is not enough; and another is the lack of a convenient approach to analyze the protein-DNA interactions. The statistics of known protein sequences and spatial structures obtained using the PFAM database (Finn et al., 2008; Punta et al., 2012) are presented in Table 1. We can see now that the number of structures for the Zn-Cys2His2 type is large enough, although the number of complexes with DNA is much smaller. In our attempt to deduce recognition rules for this type of factors, we use a new approach for the analysis of protein-DNA contacts. This approach is based on the using of the so-called *color-coding binding tables*. Previously, we used it to determine the recognition patterns for homeodomain complexes with the operator DNA (Chirgadze et al., 2012). In the present study, we apply it to the complexes of Zn-finger transcription factors and deduce common recognition rules for the Zn-Cys2Hys2 family type. We compare the result obtained with that for the homeodomain family and comment on the differences in terms of structure, function, and evolution.

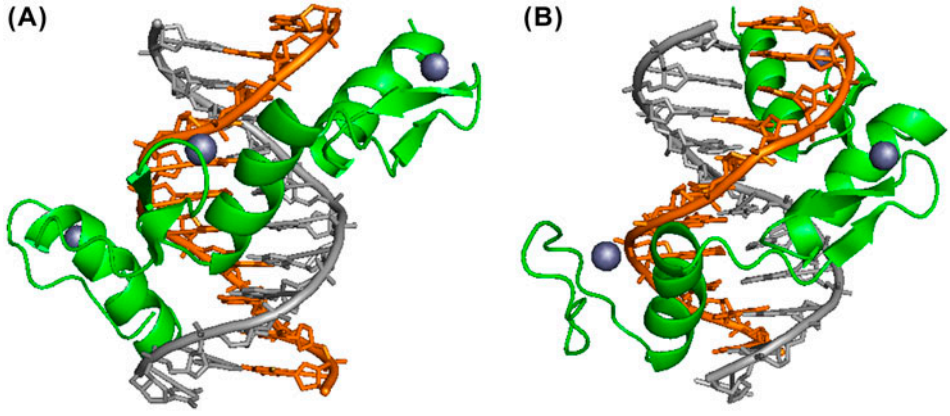

Figure 2. High resolution X-ray structure of the Zif268 complex with a fragment of operator DNA from mouse *Mus musculus* (Elrod-Erickson et al., 1996; PDB code 1aay). Views (A) and (B) differ by rotation about the vertical axis at 90 degrees. The complex contains three nucleotide triplet units and three Zn-finger protein units. The coding DNA chain is painted in orange color and ion Zn is shown as a large gray ball.

Table 1. Known sequences and spatial structures of Zn-finger transcription factors taken from the PFAM database.

PFAM accession	ID of factor types	Description	Sequences	Structures
PF09329	zf-primase	Primase zinc finger	189	0
PF03119	DNA_ligase_ZBD	NAD-dependent DNA ligase C4 zinc finger domain	2651	1
PF08996	zf-DNA_Pol	DNA polymerase alpha zinc finger	219	3
PF01258	zf-dskA_traR	Prokaryotic dksA/traR C4-type zinc finger	3394	12
PF12874	zf-met	Zinc-finger of C2H2 type	2878	70

Data and methods

The structural data-set of analyzed complexes of the Zn-Cys2His2 family of transcription factors with fragments of operator DNA comprised 22 suitable complexes, which includes 20 X-ray and two NMR structures (Table 2). That was a full number of complex structures from the 70 ones listed in the PFAM database (Punta et al., 2012) and which is presented in Table 1. Structures of most complexes were solved with high resolution of 1.5–2.6 Å, and their atomic coordinate files were taken from the Protein Data Bank (Berman et al., 2000, 2002). The various Zn-finger complexes included DNA stretches from 2 to 9 nucleotide triplets, which formed contacts with 2–9 Zn-protein units, respectively. Thus, we analyzed a total of 46 unit complexes taken from different individual Zn-factors with corresponding DNA triplets. These complexes encompassed different taxonomic groups: bacteria, insects, and mammalians, including humans. In all complexes, the recognition of DNA was defined mainly by the specific contacts of the recognizing α-helix of protein factor. The completeness of the final list of complexes was checked with the PFAM database (Punta et al., 2012), the protein-DNA interface database (Norambuena & Melo, 2010), the biomolecular interactions server IBIS (Shoemaker et al., 2012), and the BAINT database for base-amino acid interactions (Nakama, Kubota, & Sarai, 1998).

We used *the internal numbering systems* for sequences of recognizing α-helices and two DNA chains (Chirgadze et al., 2009, 2012). The first position in this system designates the first contacting amino acid residue or nucleotide. This numbering system is used in all tables and figures.

For the analysis of protein-DNA contacts, we used a new approach which is based on the so-called *color-coding binding tables*. First we calculated all interatomic protein–DNA contact distances between the atoms of different atomic groups, such as phosphate-sugar groups,

Table 2. Structural data set of complexes of Zn-Cys2His2 transcription factors with a fragment of operator DNA.

PDB code	Transcription factor	Resolution, Å	Source: common and scientific names	
1aay A	Zif268 transcription factor	1.6	House mouse	*Mus musculus*
1a1f A	Zif268 variant, GACC site	2.1	House mouse	*Mus musculus*
1a1g A	Zif268 variant, GCGT site	1.9	House mouse	*Mus musculus*
1a1h A	Zif268 variant, GCAC site	1.6	House mouse	*Mus musculus*
1a1i A	Zif268 variant, GCAC site	1.6	House mouse	*Mus musculus*
1a1j A	Zif268 variant, GCAC site	2.0	House mouse	*Mus musculus*
1a1k A	Zif268 variant, GCAC site	1.9	House mouse	*Mus musculus*
1a1l A	Zif268 variant, GCAC site	2.3	House mouse	*Mus musculus*
1f2d C	Zif268 – TATA box	2.2	House mouse	*Mus musculus*
1g2f C	Zif268 – TATA box	2.0	House mouse	*Mus musculus*
1jk1 A	Zif268 variant D20A, GCC	1.9	House mouse	*Mus musculus*
1jk2 A	Zif268 variant D20A, GCT	1.6	House mouse	*Mus musculus*
1llm D	Zif23-GCN4 chimera	1.5	House mouse	*Mus musculus*
			Baker's yeast	*Saccharomyces cerevisiae*
2wbs A	ZnF Krueppel-like factor 4	1.7	House mouse	*Mus musculus*
2i13 A	ZnF Aart (A-rich artificial)	2.0	House mouse	*Mus musculus*
2kmk A	ZnF Gfi-1 nuclear repressor	NMR	Norway rat	*Rattus norvegicus*
1ubd C	ZnF GLI initiator of mRNA	2.5	Human	*Homo sapiens*
2gli A	ZnF GLI oncogene TF	2.6	Human	*Homo sapiens*
2prt A	ZnF Wilms tumor suppressor	3.1	Human	*Homo sapiens*
2drp A	ZnF tramtrack protein	2.0	Fruit fly	*Drosophila melanogaster*
1tf3 A	ZnF factor TF IIIA	NMR	Frog	*Xenopus laevis*
1tf6 A	ZnF factor TF IIIA	3.1	Frog	*Xenopus laevis*

bases of nucleotides, and side-charged groups of amino acid residues (Chirgadze et al., 2012). We confirmed that it makes no difference with respect to the final result if we considered contacts between atoms or between atomic groups. We found that in most cases protein–DNA contacts between polar atoms were 5–6 times more frequent than contacts between the nonpolar atoms, and very often these contacts were absent at all. Therefore, we considered the contacts between atomic groups of two chains of operator DNA and the atomic groups of Zn-protein units. Further we divided all contacts into two main types depending on whether the protein contacts were with phosphate and sugar groups or with the bases of nucleotides. Then the color code was used for distinguishing these different contact types in the tables. We used yellow color for contacts of amino acid with phosphates, and cyan for contacts of amino acids with bases. The total amount of considered contacts between these groups is about an order less as compared to that between atoms. It is extremely convenient because this simplifies the analyses essentially. The number of binding contacts was estimated simply by the *frequency of occurrence* taken from the amount of specific contact observations against the total amount of observations.

We found that water-mediated contacts were insignificant for the specific protein–DNA binding, with water-mediated contacts occurring around the interfaces. Therefore, we performed the analysis of direct atomic contacts between polar atoms, and between non-polar atoms, with a proper distance threshold. In order to identify the contact type, such as polar-polar or nonpolar-nonpolar ones, we used different distance thresholds. The distance upper limits for atomic contacts were taken from the publication by Jones, Heyningen, Berman, and Thornton (1999). The direct contacts between polar atoms determined at distances less than 3.35 Å were assigned to hydrophilic interactions and could be related to hydrogen and partially to ionic bonds. Contacts between nonpolar atoms were determined at distances less than 3.9 Å. Polar-polar and nonpolar-nonpolar interatomic interactions were analyzed between the binding part of protein domains and the double-stranded DNA fragment in the region of the major groove. Contacts of the same type were defined as *invariant* if their frequencies of occurrence in the whole data-set were 80% or higher. The contacts with less than 80% frequency of occurrence were assigned to be *variable*.

The analyzed complexes of Zn-factors are different in size, which include from two to nine Zn-protein units and display some individual features. Below we describe the peculiarity of some complexes. The first studied complex is the classic complex Zif268 from mouse *Mus musculus* (Elrod-Erickson et al., 1996). Details of this structure with a very important biological function can be seen in Figure 2. Especially informative for us were 11 variants of this original complex (Elrod-Erickson, Benson, & Pabo, 1998). Here, specific residue substitutions were done in the recognizing protein helix; the others were related to differences in the DNA triplets; or both.

The human transcription factor of activating glioblastoma disease contains five Zn-finger motifs but only the second to fifth Zn-fingers are bound to DNA (Pavletich & Pabo, 1993). The fragment of operator DNA includes 20 base pairs and the DNA structure is intermediate between the structures expected for the B- and A-forms. Common protein recognizing helix binding positions 1, 2, 3, 6, and 7 show rather weak conservation of amino acids, although position 7 of histidine is always conservative.

Factor TFIIIA from *Xenopus laevis* oocytes was the first cellular gene-specific transcription factor identified in eukaryotes (Engelke, Ng, Shastry, & Roeder, 1980). It regulates the transcription of the 5S ribosomal RNA gene by RNA polymerase III binding specifically to the internal control region within the 5S RNA gene. Later, the spatial structures of various complexes of this factor with DNA were obtained. For example, factor with the PDB code 1tf3 contains three fingers f1, f2, and f3 (Wuttke, Foster, Case, Gottesfeld, & Wright, 1997). These fingers correspond to the same N-terminal part of more extended factor 1tf6 consisting of six Zn-fingers (Nolte, Conlin, Harrison, & Brown, 1998). In fact, the present model of factor TFIIIA from *Xenopus laevis* consists of nine fingers. Complexes 1tf3 and 1tf6 display somewhat different modes of binding with double-chained DNA. In particular, finger f1 from 1tf3 forms contacts with DNA, but finger f1 from 1tf6 does not form any contacts, possibly as a result of the end effect of a short fragment of DNA. Because of this, we have considered contacts of finger f1 separately for factor 1tf3 and contacts of the other fingers for factor 1tf6.

Results

Recognition motif of Zn-Cys2His2 factors

The binding of the Zn-finger factor with double-chained DNA is determined by several residues of the recognizing α-helix, which is located in the major groove of the DNA molecule. Similarly to a homeodomain, the recognizing α-helix of the Zn-finger factor also comprises 12 residues. It forms several contacts with six to nine nucleotide pairs, and this number is comparable with that for a homeodomain. However, the recognition pattern of the Zn-finger type is drastically different from the recognition patterns of a homeodomain. The main difference is that the complex of a Zn-finger works according to *the multiple module principle,* which is illustrated in Figure 3. As seen from this diagram, the DNA

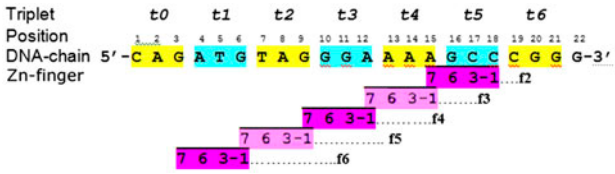

Figure 3. Schematic diagram of the modular principle of recognition of complexes of the Zn-Cys2Hys2 family with coding DNA chain. The DNA recognition site consists of consecutive nucleotide triplets. Each Zn-finger contacts DNA by four residues enumerated according to the internal numbering of the recognizing α-helix.

recognition site is constructed from consecutive triplets. It is important to note that each triplet can perform a different function. Each recognition site is described by several contacts with the nucleotide tetrad. This tetrad includes the basic nucleotide triplet and one nucleotide of the preceding triplet.

Sequence homology of recognition elements of protein and DNA sequences

Sequence alignments between the recognizing DNA tetrads, taken from total 46 complex units, are presented in Table 3. The right-hand side of the table shows the amino acid sequence alignments of the Zn-finger factor of recognizing α-helices. Note here, native Zif268 and its several mutants (var1) are also presented. As can be seen at the left-hand side of the table, guanine nucleotide is predominant in the tetrad sequence GGGG with the frequencies of occurrence 52, 54, 22, and 41%, respectively.

The protein Zn-finger factor unit is considered as a Zn-ββα protein fragment, the residues of which contribute to recognition. There are five rather conservative positions with high sequence identity of more than 70%. They include Phe (−3β), Leu (4α), His (7α), His (11α), and Thr (12α) with high sequence identities of 85, 80, 100, 87, and 70%, respectively. Surprisingly, there is only one invariant helix residue His7 (sequence identity 100%), which is directly involved in the recognition. The others seem to contribute to the stability of the protein unit Zn-ββα and Zn coordination but are believed to have no relation to the recognition. They also provide for the formation of the hydrophobic core of each protein unit.

Contacts of recognizing α-helix of Zn-finger with operator DNA

We consider the contacts between the recognizing α-helix of the protein factor and two strands of operator DNA in its major groove. The majority of protein contacts are formed with the coding DNA strand (Table 4). Each contact between nucleotide and amino acid residue is color-coded in the following way. Contacts between amino acids and phosphates are shown in yellow, and contacts between amino acids and bases in cyan. We have observed that the protein factor always binds to the tetranucleotide fragment. There are rather high total amounts of protein contacts with the DNA nucleotide sequence ZXYZ; the frequencies of occurrence are 83, 74, 67, and 61% (Table 4). The preceding nucleotide Z, which is in 52% of cases represented by guanine, forms a contact with the protein through its phosphate group (yellow color) while all nucleotides of triplet XYZ mostly form contacts with the protein through their bases (cyan color).

The coding DNA chain forms four common contacts with the protein Zn-finger factor (Table 4). In a simple form, these contacts of amino acids with nucleotide fragment $X_0Y_0Z_0X_1Y_1Z_1$ are described as follows:

Coding DNA	Protein factor	Contacts, %
Z_1	X (−1β)	76
Y_1	X (3α)	63
X_1	Arg (6α)	59
Z_0	His (7α)	83

where X is any of the amino acid residues, and the amount of contact is estimated as frequency of occurrence value, in percentage.

Four canonical contacts of the recognizing helix (−1, 3, 6, and 7) were observed in the majority of considered unit complexes (Table 4). However, only contact His7 can be considered as invariant for the whole Zn-Cys2His2 family. The unique contact was observed in 83% of complexes. It is important to note that His7 is bound with nucleotide Z preceding the basic nucleotide triplet unit. Such a contact provides for overlapping interaction with the triplets (Figure 3).

We also selected another subfamily Zn-Cys2His2-Arg, which contains Arg in position 6α. This subfamily contains 21 Zn-finger units found among the total 46 considered. The definition of the newly identified subfamily Zn-Cys2His2-Arg is 100% sequence identities of Arg6 and His7 residues. For 21 units of the Zn-Cys2His2-Arg subfamily, we have revealed the following statistics:

amino acid sequence identities: Arg6, His7 – 100 and 100% (Table 3),
frequencies of contact occurrence: Arg6, His7 – 100 and 86% (Table 4).

Three Zif268 complexes and nine Zif268 variants can be also assigned to the subfamily Zn-Cys2His2-Arg. These complexes show two invariant contact groups

Table 3. Sequence identity of recognizable nucleotide triplets of operator DNA and recognizing helix of the Zn-Cys2His2 family of transcription factors.

		DNA-contact positions of recognizing α-helix- red box																	
Complex	DNA	Amino acid position																	
PDB code	tripet	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	12	
		β	β	β	β	t	α	α	α	α	α	α	α	α	α	α	α	α	
Zif268	* *								*		*			* *					
1aay Af1	G G C G	R	R	F	S	R	S	D	E	L	T	R	H	I	R	I	H	T	
1aay Af2	G T G G	R	N	F	S	R	S	D	H	L	T	T	H	I	R	T	H	T	
1aay Af3	G G C G	R	K	F	A	R	S	D	E	R	K	R	H	T	K	I	H	L	
Zif268 var1																			
1a1k Af1	G G A C	R	R	F	S	R	S	A	D	L	T	R	H	I	R	I	H	T	
1a1f Af1	G G A C	R	R	F	S	D	S	S	N	L	T	R	H	I	R	I	H	T	
1a1g Af1	G G C G	R	R	F	S	D	S	S	N	L	T	R	H	I	R	I	H	T	
1a1h Af1	G G C A	R	R	S	F	Q	S	G	S	L	T	R	H	I	R	I	H	T	
1a1i Af1	G G C A	R	R	F	S	R	S	A	D	L	T	R	H	I	R	I	H	T	
1a1l Af1	G G C A	R	R	F	S	R	S	D	E	L	T	R	H	I	R	I	H	T	
1a1j Af1	G G C G	R	R	F	S	R	S	A	D	L	T	R	H	I	R	I	H	T	
Zif268 TATA box																			
1g2d Cf1	A A A A	R	R	F	S	Q	K	T	N	L	D	T	H	I	R	I	H	T	
1g2d Cf2	T A T A	R	N	F	S	Q	H	T	G	L	N	Q	H	I	R	T	H	T	
1g2d Cf3	C G C T	R	K	F	A	T	L	H	T	R	D	R	H	T	K	I	H	L	
1g2f Cf2	T A T A	R	N	F	S	Q	Q	A	S	L	N	A	H	I	R	T	H	T	
Zif268 var2																			
1jk1 Af1	G G C G	R	R	F	S	R	S	A	E	L	T	R	H	I	R	I	H	T	
1jk2 Af1	G G C T	R	R	F	S	R	S	A	E	L	T	R	H	I	R	I	H	T	
Zif23-GCN4																			
1llm Df2	G T G G	R	N	F	S	R	S	D	H	L	T	T	H	I	R	T	H	T	
1llm Df3	C G C G	R	K	F	A	R	S	D	E	R	K	R	H	R	D	T	I	Q	
1UBD																			
1ubd Cf1	A G A C	K	M	F	R	D	N	S	A	M	R	K	H	L	H	T	H	G	
1ubd Cf2	T G G A	K	A	F	V	E	S	S	K	L	K	R	H	Q	L	V	H	T	
1ubd Cf3	A A A T	K	R	F	S	L	D	F	N	L	R	T	H	V	R	I	H	T	
1ubd Cf4	T C A A	K	K	F	A	Q	S	T	N	L	K	S	H	I	L	T	H	A	
Aart																			
2i13 Af1	C C G G	K	S	F	S	R	S	D	H	L	A	E	H	Q	R	T	H	T	
2i13 Af2	A G C C	K	S	F	S	D	K	K	D	L	T	R	H	Q	R	T	H	T	
2i13 Af3	A A A A	K	S	F	S	Q	R	A	N	L	R	A	H	Q	R	T	H	T	
2i13 Af4	C G G A	K	S	F	S	Q	L	A	H	L	R	A	H	Q	R	T	H	T	
2i13 Af5	G T A G	K	S	F	S	R	E	D	N	L	H	T	H	Q	R	T	H	T	
2i13 Af6	G A T G	K	S	F	S	R	R	D	A	L	N	V	H	Q	R	T	H	T	
2GLI																			
2gli Af2	A G A C	L	R	P	F	A	Q	Y	M	L	V	V	H	M	R	R	H	T	
2gli Af3	C C A A	K	S	Y	S	R	L	E	N	L	K	T	H	L	R	S	H	T	
2gli Af4	C A C C	K	A	F	S	N	A	S	D	R	A	K	H	Q	N	R	T	H	
2gli Af5	G A C C	K	R	Y	T	D	P	S	S	L	R	K	H	V	K	T	V	H	
2DRP																			
2drp Af1	G G A T	R	V	Y	T	H	I	S	N	F	C	R	H	Y	V	T	S	H	
2drp Af2	T A A G	K	E	F	T	R	K	D	N	M	T	A	H	V	K	I	I	H	
2WBS																			
2wbs Af2	G G C G	W	K	F	A	R	S	D	E	L	T	R	H	Y	R	K	H	T	
2wbs Af3	G T G G	D	R	A	F	R	S	D	H	L	A	L	H	M	K	R	H	F	
1TF3																			
1tf3 Af1	G A C C	A	A	Y	N	K	N	W	K	L	Q	A	H	L	S	K	H	T	
1TF6																			
1tf6 Af2	T G G G	K	G	F	T	S	L	H	H	L	T	R	H	S	L	T	H	T	
1tf6 Af3	G G A T	L	R	F	T	T	K	A	N	M	K	K	H	F	N	R	F	H	
1tf6 Af5	T G G T	K	R	F	S	L	P	S	R	L	K	R	H	E	K	V	H	A	
2PRT																			
2prt Af2	G G C G	R	R	F	S	R	S	D	Q	L	K	R	H	Q	R	R	H	T	
2prt Af3	G G G G	R	K	F	S	R	S	D	H	L	K	T	H	T	R	T	H	T	
2prt Af4	C G C G	K	K	F	A	R	S	D	E	L	V	R	H	H	N	M	H		
2KMK																			
2kmk Af3	A C T G	K	S	F	K	R	S	S	T	L	S	T	H	L	L	I	H	S	
2kmk Af4	A T C A	K	R	F	H	Q	K	S	D	M	K	K	H	T	F	I	H	T	
2kmk Af5	T A A A	K	A	F	S	Q	S	S	N	L	I	T	H	S	R	K	H	T	
Consensus	G G G G	R	R	F	S	R	S	D	N	L	T	R	H	I	R	I	H	T	
Identity,%	52 54 22 41	46	20	**85**	**61**	48	52	33	24	**80**	40	46	**100**	35	61	37	**87**	**70**	

Table 4. Binding contacts of recognizable triplets of operator DNA and recognizing protein helix in the Zn-Cys2His2 family transcription factors. Color coding of contact types: yellow – amino acids with phosphates, and cyan – amino acids with bases.

Coding DNA chain

Complex PDB code	DNA tripet underlined	Amino acid position -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12
		β β β t α α α α α α α α α α α
Zif268	* * * *	* * * *
1aay Af1	G G C G	R R F S R S D E L T R H I R I H T
1aay Af2	G T G G	R N F S R S D H L T T H I R T H T
1aay Af3	G G C G	R K F A R S D E R K R H T K I H L
Zif268 var1		
1a1k Af1	G G A C	R R F S R S A D L T R H I R I H T
1a1f Af1	G G A C	R R F S D S S N L T R H I R I H T
1a1g Af1	G G C G	R R F S D S S N L T R H I R I H T
1a1h Af1	G G C A	R R S F Q S G S L T R H I R I H T
1a1i Af1	G G C A	R R F S R S A D L T R H I R I H T
1a1l Af1	G G C A	R R F S R S D E L T R H I R I H T
1a1j Af1	G G C G	R R F S R S A D L T R H I R I H T
Zif268 TATA box		
1g2d Cf1	A A A A	R R F S Q K T N L D T H I R I H T
1g2d Cf2	T A T A	R N F S Q H T G L N Q H I R T H T
1g2d Cf3	C G C T	R K F A T L H T R D R H T K I H L
1g2f Cf2	T A T A	R N F S Q Q A S L N A H I R T H T
Zif268 var2		
1jk1 Af1	G G C G	R R F S R S A E L T R H I R I H T
1jk2 Af1	G G C T	R R F S R S A E L T R H I R I H T
Zif23-GCN4		
1llm Df2	G T G G	R N F S R S D H L T T H I R T H T
1llm Df3	C G C G	R K F A R S D E R K R H R D T I Q
1UBD		
1ubd Cf1	A G A C	K M F R D N S A M R K H L H T H G
1ubd Cf2	T G G A	K A F V E S S K L K R H Q L V H T
1ubd Cf3	A A A T	K R F S L D F N L R T H V R I H T
1ubd Cf4	T C A A	K K F A Q S T N L K S H I L T H A
Aart		
2i13 Af1	C C G G	K S F S R S D H L A E H Q R T H T
2i13 Af2	A G C C	K S F S D K K D L T R H Q R T H T
2i13 Af3	A A A A	K S F S Q R A N L R A H Q R T H T
2i13 Af4	G G G A	K S F S Q L A H L R A H Q R T H T
2i13 Af5	G T A G	K S F S R E D N L H T H Q R T H T
2i13 Af6	G A T G	K S F S R R D A L N V H Q R T H T
2GLI		
2gli Af2	A G A C	L R P F A Q Y M L V V H M R R H T
2gli Af3	C C A A	K S Y S R L E N L K T H L R S H T
2gli Af4	C A C C	K A F S N A S D R A K H Q N R T H
2gli Af5	G A C C	K R Y T D P S S L R K H V K T V H
2DRP		
2drp Af1	G G A T	R V Y T H I S N F C R H Y V T S H
2drp Af2	T A A G	K E F T R K D N M T A H V K I I H
2WBS		
2wbs Af2	G G C G	W K F A R S D E L T R H Y R K H T
2wbs Af3	G T G G	D R A F R S D H L A L H M K R H F
1TF3		
1tf3 Af1	G A C C	A A Y N K N W K L Q A H L S K H T
1TF6		
1tf6 Af2	T G G G	K G F T S L H H L T R H S L T H T
1tf6 Af3	G G A T	L R F T T K A N M K K H F N R F H
1tf6 Af5	T G G T	K R F S L P S R L K R H E K V H A
2PRT		
2prt Af2	G G C G	R R F S R S D Q L K R H Q R R H T
2prt Af3	G G G G	R K F S R S D H L K T H T R T H T
2prt Af4	C G C G	K K F A R S D E L V R H H N M H
2KMK		
2kmk Af3	A C T G	K S F K R S S T L S T H L L I H S
2kmk Af4	A T C A	K R F H Q K S D M K K H T F I H T
2kmk Af5	T A A A	K A F S Q S S N L I T H S R K H T
Register		-5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12
Consensus	X X X X	- - - - X - - X - - R H - - - - -
Contacts	38 34 31 28	35 29 21 38 of total 46
Occurrence, %	83 74 67 61	76 63 55 83 of total 100%

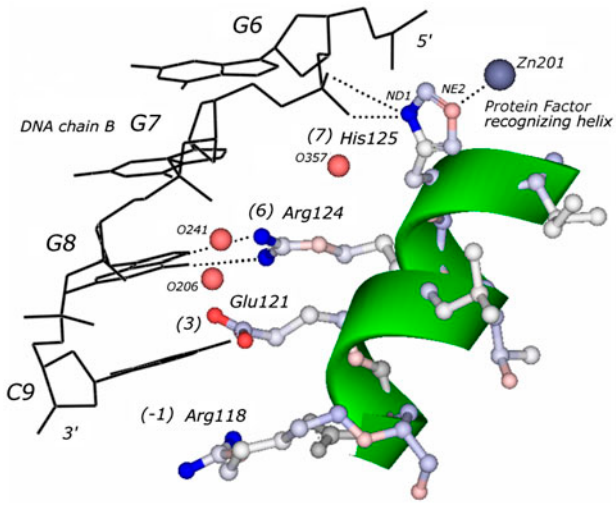

Figure 4. Two conservative invariant protein–DNA contacts in the complex of transcription factor Zif268 from mouse *Mus musculus* (PDB code 1aay). Oxygen atoms of water molecules are shown as red balls.

triplet-t1 (G8C9G10): base G8 …… NH1, NH2 Arg α6
triplet-t2 (T5G6G7): phos G7 …… ND1 His α7.

These contact groups are revealed in the structure of complex Zif268 from mouse *Mus musculus* in Figure 4 (PDB code 1aay). Here, Arg124 occupies helix position α6 and His125 – position α7. Arginine 124 side groups NH1 and NH2 form two contacts with O6 and N7 of guanine G8. Histidine side group ND1 forms two bonds with phosphate OP1 and OP2. As seen from this figure, His125 binds group NE2 via the Zn201 ion. Three water molecules, situated nearby, do not form any hydrogen bonds with these residues and are not involved in the protein–DNA binding.

For the non-coding DNA strand, we observed only a single variable contact at helical position 2 in half of all cases (Table 5). This contact is formed alternatively with the bases of any position of triplet nucleotide.

Recognition rules of Zn-Cys2His2 transcription factor with operator DNA

Initially, our analysis was carried out in the atomic approximation by calculating the interatomic contact distances. Here, we present the results in a simple diagram showing protein–DNA contacts between their atomic groups. For example, the contacts with nucleotide bases imply interactions with their groups. Structural diagrams of recognition patterns of DNA and protein helix for the Zn-Cys2His2 family are presented in Figure 5. Contact patterns for DNA and the protein recognizing helix are presented separately. All contacts were divided into two groups: invariant contacts are shown in black color and variable contacts in grey. Non-contacting atomic groups are shown as white boxes. There is only one invariant phosphate contact in the coding DNA chain, which is formed with His7 of the recognizing helix. In contrast to the recognizing region of the homeodomain factor, which forms contacts of the recognizing helix with about seven base pairs (Chirgadze et al., 2012), here we observed much fewer contacts only at the N-terminal part of the recognizing helix. Note that we consider the complex of a single nucleotide triplet, but, in fact, the Zn-finger factors contain at least two such modules or even more.

Structural diagrams of the recognition pattern for complexes of the Zn-Cys2His2 factor family and Zn-Cys2His2-Arg subfamily are shown in Figure 6. Most significant invariant contacts, composing the recognition rule, are highlighted in pink color. For the Zn-Cys2His2 family a single contact of His7 with the phosphate group of the DNA chain is observed in 83% of the cases. For the Zn-Cys2His2-Arg subfamily, we observed two contacts of Arg6 and His7 with the bases and phosphate groups of the coding DNA, which have frequencies of occurrence of 100 and 90%, for both contacts. This allows us to formulate the following *recognition rules*:

For the Zn-Cys2His2 family, there is one binding contact:

ND1 His (7) : Phos Z^{coding} (−1)

For the Zn-Cys2His2-Arg subfamily, there are two binding contacts:

ND1 His (7α) : Phos Z^{coding} (−1)
NH1, NH2 Arg (6α) : Base X^{coding} (1),

where (−1) indicates the position Z of the triplet preceding the given triplet XYZ.

Discussion

We have deduced here recognition rules for Zn-finger transcription factors in the complexes with operator DNA. The general principle of binding is a modular system of Zn-Cys2His2 complexes as shown in Figure 3. One basic nucleotide triplet is recognized by the recognizing α-helix of a unit of Zn-finger domain. Amino acids at positions −1, 3, and 6 of this α-helix recognize the nucleotides at positions Z, Y, and X of the coding DNA chain, respectively, as shown in Figure 4. The majority of these contacts are formed with the bases of the nucleic acids. The most significant result is the revealing of general recognition rules both for the

Table 5. Binding contacts of recognizable triplets of operator DNA and recognizing protein helix in the Zn-Cys2His2 family transcription factors. Contact types are color marked: yellow – amino acids with phosphates, and cyan – amino acids with bases.

Non-coding DNA chain

Complex PDB code	DNA triplet underlined	Amino acid position -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12
		β β β t α α α α α α α α α α α α
Zif268		* *
1aay Af1	C C G C A	R R F S R S D E L T R H I R I H T
1aay Af2	C A C C C	R N F S R S D H L T T H I R T H T
1aay Af3	T C G C A	R K F A R S D E R K R H T K I H L
Zif268 var1		
1a1k Af1	C C T G G	R R F S R S A D L T R H I R I H T
1a1f Af1	C C T G G	R R F S D S S N L T R H I R I H T
1a1g Af1	C C G C A	R R F S D S S N L T R H I R I H T
1a1h Af1	C C G T G	R R S F Q S G S L T R H I R I H T
1a1i Af1	C C G T G	R R F S R S A D L T R H I R I H T
1a1l Af1	C C G T G	R R F S R S D E L T R H I R I H T
1a1j Af1	C C G T A	R R F S R S A D L T R H I R I H T
TATAbox		
1g2d Cf1	T T T T C	R R F S Q K T N L D T H I R I H T
1g2d Cf2	A T A T T	R N F S Q H T G L N Q H I R T H T
1g2d Cf3	G C G A T	R K F A T L H T R D R H T K I H L
1g2f Cf2	A T A T T	R N F S Q Q A S L N A H I R T H T
Zif268 var2		
1jk1 Af1	C C G C T	R R F S R S A E L T R H I R I H T
1jk2 Af1	C C G A C	R R F S R S A E L T R H I R I H T
Zif23-GCN4		
1llm Df2	C A C C C	R N F S R S D H L T T H I R T H T
1llm Df3	G C G C A	R K F A R S D E R K R H R D T I Q
1UBD		
1ubd Cf1	T C T G G	K M F R D N S A M R K H L H T H G
1ubd Cf2	A C C T C	K A F V E S S K L K R H Q L V H T
1ubd Cf3	T T T A C	K R F S L D F N L R T H V R I H T
1ubd Cf4	A G T T T	K K F A Q S T N L K S H I L T H A
ZnF Aart		
2i13 Af1	G G C C C	K S F S R S D H L A E H Q R T H T
2i13 Af2	T C G G G	K S F S D K K D L T R H Q R T H T
2i13 Af3	T T T T C	K S F S Q R A N L R A H Q R T H T
2i13 Af4	C C C T T	K S F S Q L A H L R A H Q R T H T
2i13 Af5	C A T C C	K S F S R E D N L H T H Q R T H T
2i13 Af6	C T A C A	K S F S R R D A L N V H Q R T H T
2GLI		
2gli Af2	T C T G C	L R P F A Q Y M L V V H M R R H T
2gli Af3	G G T T C	K S Y S R L E N L K T H L R S H T
2gli Af4	G T G G G	K A F S N A S D R A K H Q N R T H
2gli Af5	T G G T G	K R Y T D P S S L R K H V K T V H
2DRP		
2drp Af1	C C T A T	R V Y T H I S N F C R H Y V T S H
2drp Af2	A T T C C	K E F T R K D N M T A H V K I I H
2WBS		
2wbu Af2	C C G C A	W K F A R S D E L T R H Y R K H T
2wbu Af3	C T C C G	D R A F R S D H L A L H M K R H F
1TF3		
1tf3 Af1	C T C T G	A A Y N K N W K L Q A H L S K H T
1TF6		
1tf6 Af2	A C C C T	K G F T S L H H L T R H S L T H T
1tf6 Af3	C C T A C	L R F T T K A N M K K H F N R F H
1tf6 Af5	A C C A A	K R F S L P S R L K R H E K V H A
2PRT		
2prt Af2	C A G A C	R R F S R S D Q L K R H Q R R H T
2prt Af3	A C G C C	R K F S R S D H L K T H T R T H T
2prt Af4	C C C C C	K K F A R S D E L V R H H N M H
2KMK		
2kmk Af3	T G A C G	K S F K R S S T L S T H L L I H S
2kmk Af4	T A G T G	K R F H Q K S D M K K H T F I H T
2kmk Af5	A T T T A	K A F S Q S S N L I T H S R K H T
Register		-5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11 12
Consensus	- - - - -	- - - - - - X - - - - - - - - -
Contacts		23 of total 46
Occurrence, %		50 of total 100%

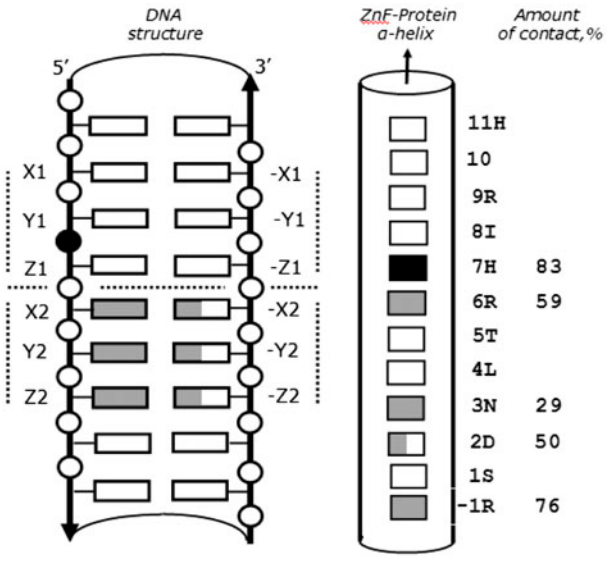

Figure 5. Recognition patterns of the operator DNA with two triplets (left) and recognizing α-helix of the Zn-Cys2His2 factor (right). All contacting groups of DNA and the recognizing protein α-helix are colored: invariant contacts are in black, variable contacts are in gray, and non-binding groups are in white. An alternatively bound helix amino acid residue at position 2 is shown as a gray–white box.

Zn-Cys2His2 family and Zn-Cys2His2-Arg subfamily (Figure 5).

The arginine Arg6α, preceding the family-defining histidine His7α, recognizes the guanine in a highly specific position. Moreover, if such an arginine is missing, its 'guanine-recognizing' role is played by its neighbor downwards along the recognizing helix (the nearest located). That can be either asparagine or histidine (both base-binding), as seen in Table 4. Therefore, the above subfamily is identified by double, both protein and DNA, sequence-specific recognition of guanine by arginine or histidine as prescribed by the original protein-DNA recognition code (Choo & Klug, 1997), but applied in a (doubly) sequence-specific manner. Therefore, alternatively the subfamily Zn-Cys2His2-Arg can well be defined as the 'sequence-specific guanine-recognizing C2H2' or even 'Klug-code C2H2' zinc fingers.

However, the most interesting feature of the interface in the Zn-Cys2His2 – DNA complex is the invariant contact of His (7α) with the coding DNA chain:

$$\text{ND1 His (7}\alpha\text{)} : \text{Phos } Z^{\text{coding}} (-1).$$

Here the phosphate group of nucleotide Z at the position −1, preceding the position of each triplet of the coding chain, is recognized by histidine 7 of the recognizing α-helix. Very specific features of this contact are as follows. Firstly, histidine 7 is involved in ligating the zinc ion. Secondly, this amino acid is absolutely (100%) conservative unlike the other histidine at positions 11 of the recognizing helix, which is also involved in zinc ligation (Table 3). Thirdly, this histidine-phosphate contact is observed in 83% of the considered cases. Finally, it is always present in at least one of the domains of the analyzed Zn-finger-DNA complexes.

It is important to compare the binding features of DNA interfaces of a homeodomain and Zn-Cys2His2 complexes. Sequences of the recognizing nucleotide triplets in Zn-finger complexes show that guanines prevail at positions −1 and 1 and are strongly represented in position 3. However, in the homeodomain complex, no 'canonical' DNA sequence motif analogous to that has been found. It suggests the diversity of Zn-finger factors is largely based on the combination of several Zn-protein

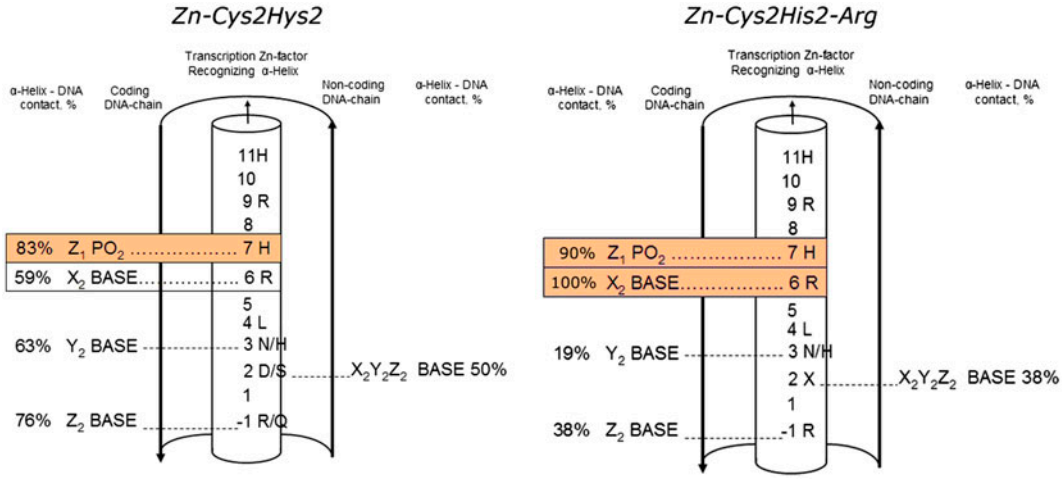

Figure 6. General diagrams of the recognition pattern for complexes of the Zn-Cys2His2 family and Zn-Cys2His2-Arg subfamily transcription factor with operator DNA. Most significant invariant contacts with higher frequency occurrences, consisting the recognition rule, are marked in pink color.

Table 6. Binding features of complexes of operator DNA with transcription factors for the homeodomain and Zn-Cys2His2 families.

Feature	Homeodomains	Zn-Cys2His2 Family
DNA molecule	Double-chained helix	Double-chained helix
DNA binding site	Major grove of the molecule	Major grove of the molecule
DNA recognition site	Total 7–8 nucleotide pairs for one factor	Total 6–27 nucleotide pairs, one factor includes a few triplet units
DNA sequence site	XYZ-TAAT-XYZ	XYZ-XYZ – any nucleotide
Protein molecule	Protein 3α peptide unit	Zn-finger 2βα peptide unit
Protein binding site	Recognizing α-helix total 12 residues	Recognizing α-helix – 12 residues and β-strand – 1 residue
Protein site sequence	10 residues of total 12	5 residues of total 13
Invariant contacts	Two contacts – coding DNA chain One contact – non-coding DNA chain	One contact – coding DNA chain
Variable contacts	Two contacts – coding DNA chain Two contacts with non-coding DNA chain	Three contacts – coding DNA chain One contact with non-coding DNA chain
Recognition rules	One contact Asn-Water-Ade Six contacts of α-helix with phosphate of DNA	One contact of α-helix His7 phosphates of DNA

units. The structural arrangement of such complexes also supports this assumption (Figure 3). All the features of two complex types are summarized in Table 6. In both cases, the main binding site is related to the major groove of the DNA molecule but the binding site sequences differ essentially. In contrast to a homeodomain, Zn-finger factors contact mostly the coding chain, generally the bases. Such an arrangement allows the binding side chains of amino acids to be closer to the recognized bases of a single coding chain, which practically eliminates the necessity to use the other chain for recognition. As a result, the recognition rules for both complexes are very different.

We believe that the used alignment algorithm of protein–DNA complexes, specifically their interfaces, and the novel theoretical understanding of the protein–DNA binding could be also applied to protein–DNA complexes of other types. Importantly, wide-ranging sense of our conception for recognition rules could be generalized to different DNA-binding proteins.

We suggest that *invariant* residues, which form very specific contacts with the coding DNA-chain, are strongly responsible for creation of the complex of these Zn-factor classes. And, as we know, this is conditioned by strong interactions of oppositely charged electrostatic protein and DNA surfaces in the contacting regions. In contrast, the *variable* contact residues stipulate only the binding of a definite Zn-factor, which is responsible only for its specific feature. It is difficult to say now what type of contacts is more important for stability of the protein-DNA complex. But they both lead to the formation of the complex.

The multiple modular principle of the binding of Zn-finger factors to operator DNA is related to the variability of their functions (Condit & Railsback, 2007). A bright example of the complicated structure of six-finger factor TFIII with a fragment of operator DNA was presented earlier (Nolte et al., 1998). The factor TFIII can recognize different and separated DNA sequences by using many Zn-protein factor units.

The practical use of the presented results in terms of designing Zn-finger motifs with specific DNA-binding functions is strongly defined by the recognition codes formulated herein. If designed mutations eliminate the code-forming histidine-phosphate or arginine–base interactions, the binding mode of the mutated Zn-finger motif to DNA can be drastically changed. And this means that such mutations should be avoided.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2013.879074.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research; Project No. 11-07-00374a. We thank the reviewers for the valuable notes which improved the manuscript. One of us, professor Yuri N. Chirgadze, would like to express his sincere gratitude to his wife for her invaluable help and encouragement during the work on this study.

References

Andreini, C., Banci, L., Bertini, I., & Rosato, A. (2006). Zinc through the three domains of life. *Journal of Proteome Research, 5*, 3173–3178.

Auld, D. S. (2001). Zinc coordination sphere in biochemical zinc sites. *Biometals, 14*, 271–313.

Benos, P. V., Lapedes, A. S., & Stormo, G. D. (2002). Is there a code for protein DNA recognition? Probabilistically. *Bioessays, 24*, 466–475.

Berg, J. M., & Shi, Y. (1996). The galvanization of biology: A growing appreciation for the roles of zinc. *Science, 271*, 1081–1085.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., … Zardecki, C. (2002). The protein data bank. *Acta Crystallographica, D58*, 899–907.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The RCSB protein data bank. *Nucleic Acids Research, 28*, 235–242.

Chirgadze, N. Yu., Sivozhelezov, V. S., Polozov, R. V., Stepanenko, V. A., & Ivanov, V. V. (2012). Recognition rules for binding of homeodomains to operator DNA. *Journal of Biomolecular Structure & Dynamics, 29*, 715–731.

Chirgadze, N. Yu., Zheltukhin, E. I., Polozov, R. V., Sivozhelezov, V. S., & Ivanov, V. V. (2009). Binding regularities in complexes of transcription factors with operator DNA: Homeodomain family. *Journal of Biomolecular Structure & Dynamics, 26*, 687–700.

Choo, Y., & Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Current Opinion in Structural Biology, 7*, 117–125.

Christianson, D. W. (1991). Structural biology of zinc. *Advances in Protein Chemistry, 42*, 281–355.

Condit, C. M., & Railsback, L. B. (2007). Generalization through similarity: Motif discourse in the discovery and elaboration of zinc finger proteins, *BioMed Central, Journal of Biomedical Discovery and Collaboration, 2*, 5. doi:10.1186/1747-5333-2-5

Desjarlais, J. R., & Berg, J. M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proceedings of the National Academy of Sciences, 89*, 7345–7349.

Elrod-Erickson, M., Benson, T. E., & Pabo, C. O. (1998). High resolution structures of variant Zif268-DNA complexes: Implications for understanding zinc finger-DNA recognition. *Structure, 6*, 451–464.

Elrod-Erickson, M., Rould, M. A., Nekludova, L., & Pabo, C. O. (1996). Zif268 protein-DNA complexes refined at 1.6A: A model system for understanding zinc finger-DNA interactions. *Structure, 4*, 1171–1180.

Engelke, D. R., Ng, S. Y., Shastry, B. S., & Roeder, R. G. (1980). Specific interaction of a purified transcription factor with an internal control region of 5S RNA genes. *Cell, 19*, 717–728.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., … Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research, 36*, D281–D288.

Freemont, P. S., Lane, A. N., & Sanderson, M. R. (1991). Structure aspects of protein-DNA recognition. *Biochemical Journal, 278*, 1–23.

Gamsjaeger, R., Liew, C. K., Loughlin, F. E., Crossley, M., & Mackay, J. P. (2007). Sticky fingers: Zinc-fingers as protein-recognition motifs. *Trends in Biochemical Sciences, 32*, 63–70.

Jones, S., Heyningen, P., Berman, H. M., & Thornton, J. M. (1999). Protein-DNA interactions: A structural analysis. *Journal of Molecular Biology, 287*, 877–896.

Klug, A. (2010). The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annual Review of Biochemistry, 79*, 213–231.

Krishna, S. S., Majumdar, I., & Grishin, N. V. (2003). Structural classification of zinc fingers; survey and summary. *Nucleic Acids Research, 31*, 532–550.

Laity, J. H., Lee, B. M., & Wright, P. E. (2001). Zinc finger proteins: New insights into structural and functional diversity. *Current Opinion in Structural Biology, 11*, 39–46.

Matthews, B. W. (1988). Protein-DNA interaction. No code for recognition. *Nature, 335*, 294–295.

McCammon, J. A. (1998). Theory of biomolecular recognition. *Current Opinion in Structural Biology, 8*, 245–249.

Miller, J., McLachlan, A. D., & Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus oocytes*. *EMBO Journal, 4*, 1609–1614.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology, 247*, 536–540.

Nakama, T., Kubota, Y., & Sarai, A. (1998). 3DinSight: An integrated relational database and search tool for the structure, function and properties of biomolecules. *Bioinformatics, 14*, 188–195.

Nolte, R. T., Conlin, R. M., Harrison, S. C., & Brown, R. S. (1998). Differing roles for zinc fingers in DNA recognition: Structure of a six-finger transcription factor IIIA complex. *Proceedings of National Academy of Sciences of USA, 17*, 2938–2943.

Norambuena, T., & Melo, F. (2010). The protein-DNA interface database. *BioMed Central, Bioinformatics, 11*, 262–274.

Paillard, G., & Lavery, R. (2004). Analyzing protein-DNA recognition mechanisms. *Structure, 12*, 113–122.

Patel, K., Kumar, A., & Durani, S. (2007). Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. *Biochimica et Biophysica Acta, Protein and Proteomics, 1774*, 1247–1253.

Pavletich, N. P., & Pabo, C. O. (1991). Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1Å. *Science, 252*, 809–817.

Pavletich, N. P., & Pabo, C. O. (1993). Crystal structure of a five – finger GLI-DNA complex: New perspectives on Zn fingers. *Science, 261*, 1701–1707.

Ponomarenko, J. V., Ponomarenko, M. P., Frolov, A. S., Vorobyev, D. G., Overton, G. C., & Kolchanov, N. A. (1999). Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics, 15*, 654–668.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families' database. *Nucleic Acids Research, 40*, D290–301.

Reddy, C. K., Das, A., & Jayaram, B. (2001). Do water molecules mediate protein-DNA recognition? *Journal of Molecular Biology, 314*, 619–632.

Rhodes, D., Schwabe, J. W., Chapman, L., & Fairall, L. (1996). Towards an understanding of protein-DNA recognition. *Philosophical Transactions of the Royal Society, London B, 351*, 501–509.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry, 79*, 233–269.

Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., … Panchenko, A. R. (2012). IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Research, 40*, D834–D840.

Steffen, N. R., Murphy, S. D., Tolleri, L., Hatfield, G. W., & Lathrop, R. H. (2002). DNA sequence and structure: Direct and indirect recognition in protein-DNA binding. *Bioinformatics, 18*, S22–S30.

Surai, A., & Kono, H. (2005). Protein-DNA recognition patterns and prediction. *Annual Review of Biophysics: Biomolecular Structures, 34*, 379–395.

Suzuki, M., Brenner, S. E., Gerstein, M., & Yagi, N. (1995). DNA recognition code of transcription factors. *Protein Engineering, 8*, 319–328.

Suzuki, M., Gerstein, M., & Yagi, N. (1994). Stereochemical basis of DNA recognition by Zn finger. *Nucleic Acids Research, 22*, 3397–3405.

Wolfe, S. A., Nekludova, L., & Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure, 29*, 183–212.

Wuttke, D. S., Foster, M. P., Case, D. A., Gottesfeld, J. M., & Wright, P. E. (1997). Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: Determinants of affinity and sequence specificity. *Journal of Molecular Biology, 17*, 183–206.