Mining Frequent Spatial Docking Patterns in Zinc Finger - DNA Complexes

Andrea Szabóová[1], Ondrej Kuželka[1], Filip Železný[1] and Jakub Tolar[2]

[1] Czech Technical University, Prague, Czech Republic
[2] University of Minnesota, Minneapolis, United States of America

Abstract— **We use techniques of logic-based relational machine learning to automatically detect spatial patterns common to 21 previously described examples of zinc finger - DNA complexes. We demonstrate that such patterns can be found and thus the proposed methodology may potentially serve to achieve better understanding of zinc finger - DNA binding.**

Keywords— **Structural Genomics, Machine Learning, Data Mining**

I. Introduction

Computational modeling of protein - DNA interactions has recently received significant attention [1, 2, 3]. Models predicting such interactions may become an important tool for hypothesizing about unknown gene regulatory pathways, to name just one example of their potential impact.

Here we specifically focus on the class of proteins known as zinc fingers (ZF) [4] which are the most common transcription factors in various organisms including humans. Zinc fingers are especially attractive due to the recent advances in generating customized *zinc finger nucleases* (ZNFs) [5]. ZFNs consist of an array of zinc fingers binding to a specific locus of a target DNA, and a nuclease inducing a double stranded break at that locus. As such, ZFNs may serve as a means for DNA editing with impact in novel gene therapy methods. When compared to other gene therapy methods, ZFNs are unique in disrupting the genomic DNA only transiently. This is critical for any clinical translation since the viruses that are inserted into genomes and used for gene therapy today have been shown to lead to significant and life-threatening side effects of their own and thus are unlikely to be widely accepted in clinical gene therapy. To achieve a seamless gene correction, ZFN technology exploits a cells own ability to repair broken DNA. When DNA breaks from exposure to its environment, DNA-repairing enzymes in the cell find and re-join the two exposed DNA ends. If another piece of DNA, such as a fully functional gene, is floating around, it can replace the broken DNA during this repair instead. In this process (called gene targeting or gene editing) the defective gene is permanently replaced by a healthy one.

To design zinc finger arrays of sufficient binding specificity, it is important to understand the patterns underlying ZF-DNA binding. The elementary binding principles have been described [6] for certain prototypical zinc fingers. No computational model has however been yet developed that would reliably predict binding for an arbitrary zinc finger domain an arbitrary DNA site. It is unlikely that such a model will be deductively inferred from biochemical laws. Current approaches thus take the opposite way, where binding rules are *learned* by generalization from sets of known ZF-DNA complexes. The study [7] adopts the assumption that in any ZF-DNA complex, three residue-nucleotide pairs can be identified that are most critical for the binding to occur. A probabilistic model called *Hidden Markov Model* is then learned, assigning a binding probability to any sequence of three such pairs. A more general strategy is followed in [8] which describes each complex by attributes corresponding to all residue-nucleotide pairs. A *support vector machine* classifier is then learned to discriminate the binding examples from the non-binding. [9] estimates the sensitivity of binding predictions to the docking geometry. Unlike the two former studies, [9] thus accounts also for the spatial structure of the complexes. However, [9] does not explicitly identify structural patterns characteristic for binding; rather, geometrical properties are only considered for assessing the similarity of pairs of ZF-DNA complexes.

In contrast to the mentioned studies, here we aim at explicitly discovering structural patterns characteristic for ZF-DNA complexes. We do not bias our approach by any prior hypothesis about ZF-DNA binding. Rather, our algorithm only receives the spatial structure of exemplary complexes, and is required to find non-trivial substructures that occur in all these input complexes, i.e. in the learning data. Here, the learning data are described by the 3D coordinates of all residues and nucleotides in the vicinity of the respective binding sites. The top row in Figure 2 is a graphical presentation of three elements drawn from the learning set. Simply put, the algorithm should answer the question what such images have in common. We however impose additional requirements to prevent uninteresting answers corresponding e.g. to obvious structures found only in one of the constituents (ZF or DNA). To implement the desired functionality we employ techniques of *logic-based relational machine learning* [10] which is a novel

branch of machine learning, concerned with the discovery of patterns in complex relational data structures.

II. Materials and Methods

Data. Twenty-one examples of Cys_2His_2 ZF-DNA complexes were sourced from [9]. Their structural description was obtained from the *Protein Data Bank*. Some of the proteins in fact contain multiple zinc finger sites with bound DNA. One option is to treat each of them as an independent complex, resulting in ninety-three individual ZF domains in complex with DNA. In this study we rather followed a more general goal of searching patterns in the entire combined complexes, possibly containing multiple ZF domains. Therefore we had twenty-one learning examples that were further arranged as follows to be processible by our pattern discovery method.

From the structural description of each complex, we extracted the following information retained for subsequent experiments i) the list of all contained nucleotides with information on their type (a,g,c,t) and the containing strand of the DNA (10's of nucleotide entries per complex), ii) the list of all contained contained residues with information on their type (10's of residue entries per complex), iii) the list of pairwise spatial distances among all nucleotides and residues (100's to 1000's entries per complex).

Method. Our pattern discovery method assumes that ZF-DNA complexes are described by means of formal-logic assertions. For example, the assertion $Nuc(\mathsf{n1},\mathsf{c},\mathsf{st1})$ denotes that the first described nucleotide (with id $\mathsf{n1}$) is a cytosin (c) on the first DNA strand ($\mathsf{st1}$). A complete description of a ZF-DNA complex is a logical conjunction of such statements (so called *literals*), pertaining to all involved nucleotides, residues, and their all pairwise spatial distances. A small fragment of such an exemplary conjunction is

$$\begin{aligned} D = & Nuc(\mathsf{n1},\mathsf{c},\mathsf{st1}) \wedge Res(\mathsf{r1},\mathsf{his}) \wedge Dist(\mathsf{n1},\mathsf{r1},16) \\ & \wedge Nuc(\mathsf{n2},\mathsf{g},\mathsf{st1}) \wedge Res(\mathsf{r2},\mathsf{arg}) \wedge \ldots \end{aligned}$$

A full description of a real ZF-DNA complex corresponds in fact to a much larger conjunction containing hundreds of literals. It would be also possible to describe complexes on a finer level of detail, where each literal would correspond to an atom or a distance between two atoms. For clarity, we follow here the higher level of abstraction as illustrated above.

A *pattern* P is also a logical formula that is *weaker* than D in the sense that whenever D is true then P must also be true, i.e. D *implies* P, which is denoted as $D \models P$. It is trivial to find a pattern in D, one such pattern is

$$P_1 = \exists x,y,z : Nuc(x,y,z)$$

simply stipulating the presence of a nucleotide of an aribtrary id and type on an arbitrary strand (hence the existentially quantified variables x,y,z). Clearly, $D \models P_1$. A slightly less trivial pattern is, for example,

$$P_2 = \exists x,y,z : Nuc(x,y,z) \wedge Res(x,\mathsf{his}) \wedge Dist(x,y,16)$$

asserting the presence of a residue histidine (his) in distance 16A from some nucleotide. P_2 is stronger than P_1 according to the implication relation $P_2 \models P_1$, which clearly holds.

Trivial (weak) patterns such as P_1 are uninteresting from the biological point of view. We rather want to search for the strongest possible patterns. At the same time, however, we want each produced pattern to be frequent, that is, to hold in as many as possible input ZF-DNA complexes. There is an obvious trade-off between a pattern's strength on one hand, and its frequency on the other hand: the stronger a pattern is, the less likely it will hold in many complexes. Following this trade-off, we request our algorithm to produce the *strongest* possible patterns that hold for *all* input complexes. Formally, we seek patterns P complying with

1. $\forall i : D_i \models P$, where D_i are descriptions of the input ZF-DNA complexes (learning data)
2. P is as strong as possible, i.e. there is no P' complying with condition 1 such that $P' \models P$

A detailed description of the computational procedures used to accomplish this task is beyond the scope of this paper. In brief, we employ techniques of logic-based relational machine learning [10] also known as *inductive logic programming*. Here, all examples D_i and all admissible patterns P_j can be organized as vertices in a graph with oriented edges corresponding to the implication relation $\models$. This graph is usually very large due to the combinatorially large number of possibilities to formulate a 'candidate' pattern. Standard graph-search techniques (such as beam search) can be used to systematically explore this graph yielding patterns complying with the above formalized conditions. In particular, we employ our recently published algorithm [11] since it can scale to rather large structures corresponding to ZF-DNA complexes, which would be prohibitively large for mainstream inductive logic programming algorithms.

III. Results

To prevent our method from reporting patterns characterstic only for one of the two complex constituents (ZFN or

DNA) we imposed an additional condition that each pattern must assume at least one nucleotide, at least one residue, and stipulate a particular spatial distance between the two. As a result we obtained about five hundred strongest patterns present in all 21 input complexes. Due to the limited space we show here only three examples of the resulting patterns. For ease of exposition we selected pattern examples where all assumed nucleotides and residues are instantiated as to their type (e.g., patterns exemplified in the previous section do not comply with this condition) whereas the specific DNA strand of the contained nucleotides is not assumed. This enables us to present the selected patterns as simple planar graphs. These are shown in Fig. 1. To allow further insight into the meaning of the patterns, we also show the substructure corresponding to the third reported pattern (right-most in Fig. 1) in particular examples of ZF-DNA complexes. We show three such examples in the bottom row of Fig. 2 though, by construction of the discovery method, the pattern-compliant structures are present in all the input complexes.

IV. Discussion and Future Work

To our best knowledge this study is the first attempt to automatically discover common docking patterns in ZF-DNA complexes: previous approaches to learn knowledge from examples of such complexes either did not take spatial structure into account or did so only for assessing similarity between the complexes. Another point distinguishing our study from previous research is that patterns are searched not in single finger domains but rather in the entire complexes possibly consisting of multiple zinc fingers. The search results indeed contain patterns that span multiple fingers as can be seen in Fig. 2. An obvious question arises whether the discovered patterns are truly characteristic for ZF-DNA binding, or they are present in the 21 learning examples only due to chance. In a subsequent provisional experiment we found that the patterns are not frequent in complexes formed by non-ZF proteins and DNA. This finding would rather suggest the former hypothesis, although a cross-validation experiment will be needed to assess the statistical reliability of the results. Only then the discovered patterns will be ready for biochemical interpretation.

Our study thus primarily contributed by showing the feasibility of the task of discovering spatial patterns in ZF-DNA complexes. Such patterns have wide potential in further refined analysis tasks. For example, in a *clustering* task one may look for spatial patterns characteristic for mutually non-overlapping subsets of the initial set of complexes. Here, the structural patterns would act as fingerprints inducing a sort of ZF-DNA taxonomy much like primary DNA data serve to automatically induce taxonomies of species in evolutionary bioinformatics. A structural classification of zinc fingers was previously proposed in [13] but that was elaborated semi-manually. From the gene-therapy application viewpoint, however, the task of most practical interest pertains to *predictive classification*. Here, for a given zinc finger, a specific DNA sequence would be predicted where the ZF is likely to bind. Also here the present methodology would be applicable for searching patterns with high predictive power to be used as features in the construction of classifiers. In all the envisioned applications of the present methodology it is likely that the current rather simple descriptions of complexes will need to be extended to account for further relevant information pertaining to structure on the atom (rather than residue/nucleotide) level, mutual bonds and elementary properties of the residues, such as polarity or hydrophobicity.

Acknowledgements. This work was supported by project ME10047 granted by the Czech Ministry of Education and the Czech Technical University internal grant #10-801940. Thanks go to András Szilágyi for valuable consultations.

References

1. Gao M., Skolnick J.. A Threading-Based Method for the Prediction of DNA-Binding Proteins with Application to the Human Genome *PLoS Computational Biology.* 2009;5.
2. Szilágyi A., Skolnick J.. Efficient Prediction of Nucleic Acid Binding Function from Low-resolution Protein Structures *Journal of Molecular Biology.* 2006;358:922–933.
3. Siggers T. W., S. A., Honig B.. Structural Alignment of ProteinDNA Interfaces: Insights into the Determinants of Binding Specificity *Journal of Molecular Biology.* 2005;345.
4. Luchi S., Kuldell N.. *Zinc Finger Proteins: From Atomic Contact to Cellular Function*. Kluwer 2005.
5. Cathomen T., Joung J. K.. Zinc-finger Nucleases: The Next Generation Emerges *Molecular Therapy.* 2008;16.
6. Wolfe S. A., Nekludova L., Pabo C. O.. DNA recognition by Cys-2-His-2 zinc finger proteins *Annu. Rev. Biophys. Biomol. Struct.* 2000;29.
7. Cho S. Y., Chung M., Park M., Park S., Lee Y. S.. ZIFIBI: Prediction of DNA binding sites for zinc finger proteins *Biochemical and Biophysical Research Communications.* 2008;369.
8. Persikov A. V., Osada R., Singh M.. Predicting DNA recognition by Cys2 His2 zinc finger proteins *Bioinformatics.* 2009;25.
9. Siggers T. W., Honig B.. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry *Nucleic Acids Research.* 2007;35.
10. De Raedt L.. *Logical and Relational Learning*. Springer 2008.
11. Kuželka O., železný F.. Block-wise construction of acyclic relational features with monotone irreducibility and relevancy properties in *ICML '09: 26th International Conference on Machine Learning* 2009.
12. Moreland J.L., A.Gramada , Buzko O.V., Zhang Qing, Bourne P.E.. The Molecular Biology Toolkit (MBT): A Modular Platform for Developing Molecular Visualization Applications *BMC Bioinformatics.* 2005;6.
13. Krishna S. S., Majumdar I., Grishin N. V.. Structural classification of zinc fingers *Nucleic Acids Research.* 2003;31.

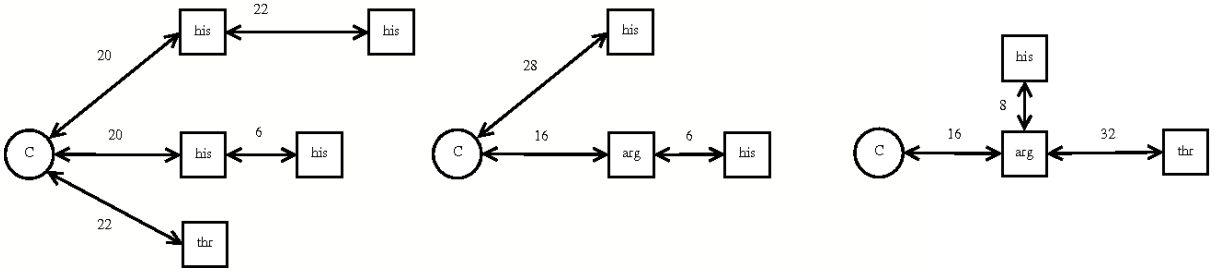

Fig. 1: Examples of patterns discovered. Edges indicate spatial distances; their lengths are shown only approximately to scale with the distances in angstroms.

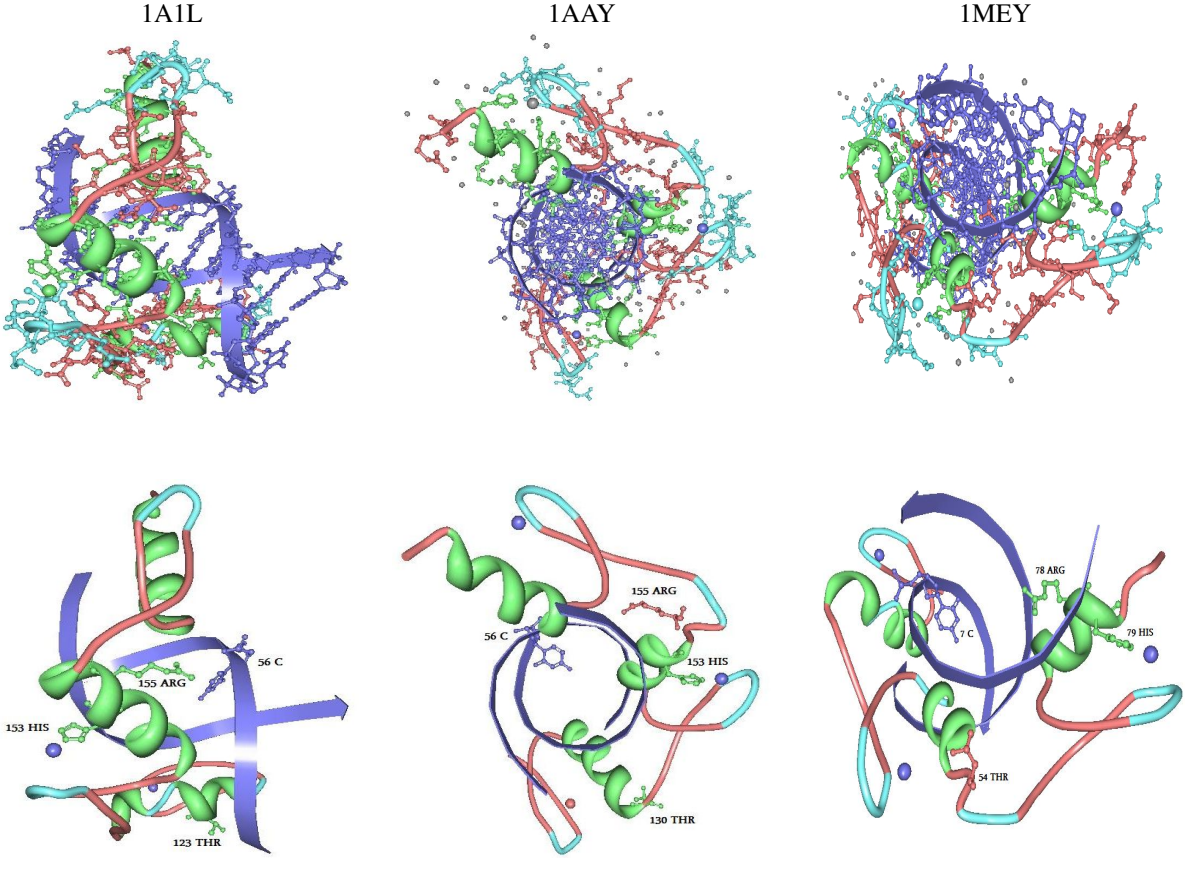

Fig. 2: Three exemplary ZF-DNA complexes (one per column) shown using the protein viewer software [12]. **Top row:** Secondary structure features (including all contained atoms and atomary bonds) are shown in red (β-sheets) and green (α-helices); the two DNA strands are shown in blue. Grey balls represent zinc atoms. **Bottom row:** Structures corresponding to one discovered pattern (the right-most in Fig. 1) shown in the same three complexes. Only those atoms from the top row are retained that correspond to the nucleotides and residues assumed by the pattern.