EFFECTS OF DISEASE-RELATED MUTATIONS ON TRANSCRIPTION FACTOR BINDING

Kirsti Laurila and Harri Lähdesmäki

Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
kirsti.laurila@tut.fi, harri.lahdesmaki@tut.fi

ABSTRACT

Many diseases are caused by hereditary mutations. So far, most of the identified mutations affect the coded protein sequence. However, an increasing number of the identified disease-related mutations occur in gene regulatory sequences. These mutations pose a threat to influence the mechanism by which a cell regulates the transcription of its genes. Here we have studied the effect of mutations on transcription factor binding affinity computationally. We have compared our results with experimentally verified cases where a mutation in the gene regulatory region either creates a new transcription factor binding site or deletes a previously existing one. We have also investigated the statistical properties of the changes on transcription factor binding affinity according to mutation type. Although accurate binding site prediction is difficult in general, our results demonstrate that computational analysis can provide valuable information about the effect of mutations on transcription factor binding sites. The analysis results also give a useful test set for the in vitro studies of regulatory mutation effects.

1. INTRODUCTION

Millions of single nucleotide polymorphisms (SNPs) are identified in the human genome. The majority of these SNPs are neutral but some of them are linked to hereditary diseases. Most of these disease-causing mutations alter the protein sequence, but a set of mutations are identified to occur in gene regulatory sequences. These mutations may cause a significant change in individuals phenotype by increasing or decreasing the gene expression levels. Some examples of this have been verified experimentally. The expression level of the gene for a 91-kD glycoprotein component of the phagocyte oxidase (gp91-*phox*) are decreased because of promoter region mutations that are are associated with X-linked chronic granulomatous disease [1]. Moreover, with Alzheimer disease patients, abnormal high expression levels of the amyloid precursor protein (APP) were measured in vitro when studying three point mutations in the APP promoter region [2].

Although all the mechanisms of gene expression regulation are not known, the mutations in the promoter regions may cause wrong transcription factor (TF) binding and this may in turn have effect on transcription levels. For instance, it has been shown experimentally that the point mutation T→C at 77 nucleotides upstream of the transcription starting site (TSS) of the δ-globin gene (HBG) changes the binding affinity of the TF GATA-1 and this also alters the expression levels of the gene. This mutation is associated with a hereditary disease δ-thalassemia [3]. Another example is described in [4] where a point mutation in 292 nucleotides upstream of the TSS of the reticulocyte-type 15-lipoxygenase-1 (ALOX15) gene causes a new transcription factor binding site (TFBS) for the TF SPI1. This again causes three-fold expression levels compared with wild type gene expression. ALOX15 has a role in the development of asthma and some other diseases.

Mutation effect on TF binding has been studied computationally in [5], where authors have used the change of a score computed based on position specific scoring matrixes (PSSMs) to infer if the binding of some TF changes. This can be problematic since a single nucleotide change usually causes a very small change in score and one cannot directly say that whether this change of score is significant or not. This fact was found when they compared the scores of mutations that are known to affect TF binding with the scores of background substitutions [5].

In this paper, we use a similar approach as in [5] to analyze the regulatory mutations and how they affect the TF binding. However, we use the p-values to compare the wildtype and mutated cases and to get the results of different genes and TFs comparable. We also study if some type of mutation is more significant than the others. This is because the DNA bending ability is known to be different for separate dinucleotide steps [6], [7]. Further, it has been found that contacts between TFs and purines are especially important and because the bending of DNA has an effect on TF binding [8], [9], [10].

2. METHODS

The mutations used in this study were the regulatory mutations from Human Gene Mutation Database (HGMD) [11]. The regulatory mutations dataset was filtered to contain only those mutations that occur upstream from transcription or translation starting sites. Altogether we used 474 mutations in 256 genes.

PSSMs are a widely used in predicting TFBSs and we

apply them in our analysis as well [12], [13]. PSSMs were collected from Transfac (Release 10.3) [14] and Jaspar [15],[16]. Only those matrixes that have been built (at least partially) using human sequences were used. After this selection, we had 496 matrixes for 343 different TFs.

The score for TF binding to the DNA sequence x_1^n was computed by

$$S(x_1^n) = \frac{P_{TF}(x_1^n)}{P_{bg}(x_1^n)}, \quad (1)$$

where $P_{TF}(x_1^n)$ is the probability computed by PSSM and $P_{bg}(x_1^n)$ is the background probability. We added a small pseudo count (0.005) to all elements in PSSM to prevent zero probabilities. As a background model, we used a third order Markov model whose parameters were computed from the promoter sequences of all human genes. As a promoter sequence, we considered commonly used 5000 bases upstream from the start of the first (according to 5' end) annotated mRNA sequence of the gene. However, promoter sequences were not allowed to overlap. The promoter sequences we used were collected from annotated sequence files (gbk-files) of human chromosomes. These files were downloaded from ftp-site of National Center for Biotechnology Information.

We computed the scores for the wildtype and the mutated sequences of our regulatory mutations dataset. Since location of mutation in putative binding sites is not known, we computed the scores for all locations within PSSM. In view of the fact that the distributions were very different for each PSSM, we did not compare the scores but computed the p-values for each mutation. To get the reference distribution for the p-value estimation, the scores were computed for each position of each promoter sequence.

Nucleotides can be divided in purines (denoted by R, consists of bases A and G) and pyrimidines (denoted by Y, bases C and T). By these classifications the dinucleotides can be divided into four classes, RR, YY, YR and RY. Further, for single point mutation the mutations can be divided into 8 groups whether the mutation is in the first or second nucleotide. We divided mutations into these classes, so that each mutation occurred both in the first and in the second nucleotide. Each mutation class was studied separately.

We made a literature search for known mutations affecting TF binding. We collected 6 experimentally proven mutations from articles and rSNP_DB [17]. These mutations were used to set a threshold for a relevant change in binding affinity.

3. RESULTS

We evaluated the effect of the experimentally verified mutations on TF binding by PSSM scores. The list of mutations and their p-values are at Table 1. All mutations showed a big change in p-value (over 0.2) between the wildtype and mutated sequence. However, the p-values of the sequence which has stronger affinity to TF were quite high in some cases i.e. the binding site was quite not statistically significant. Nevertheless, even weaker binding sites can be important, since it has been recently shown

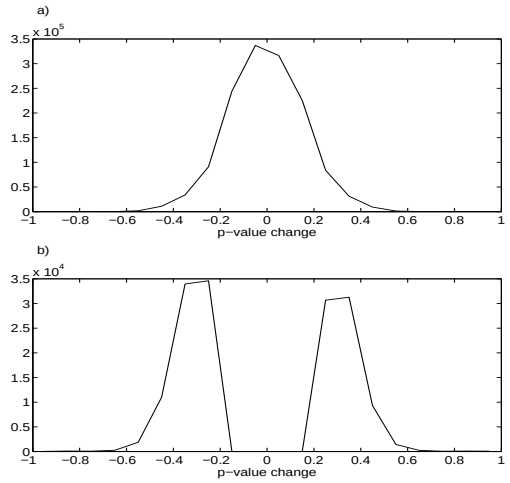

Figure 1. Distributions of the p-value changes. a) All changes. b) Only changes that exceeded the thresholds.

that models which include weak binding sites predict the expression patterns better than those models from which the weak binding sites are excluded [21].

For the experimentally verified mutations, big p-value changes are found for several PSSMs of a single TF. For example for the mutation in hemoglobin gamma G(HBG2) promoter, the p-values corresponding to 4 out of 7 PSSMs for TF SP1 showed a difference in binding affinity. However, one of the matrixes showed the change in two different matrix positions, which suggests that all of the matrixes are not very specific to the binding site.

We computed the change in p-value for each mutation (the wildtype sequence − the mutated sequence) for each TF. This p-value change was considered as a score to measure the change in TF's affinity to bind. The distribution of changes are shown in Figure 1a). Based on the experimentally verified cases we considered the change to be relevant if the p-value change (absolute value) was over 0.3 or the change was over 0.2 and p-value of either the wildtype or the mutated sequence was under 0.3. Approximately 11% of changes exceeded these boundaries. The set of experimentally verified mutations is relatively small and that prevents us from inferring more conservative thresholds without loosing too many verified cases. Current knowledge does not allow us to discriminate true and false changes more carefully (see e.g.[5]). This choice of thresholds, however, results in a set of predicted binding changes that is enriched for true binding affinity changes. Consequently, despite some false positives, our analysis results provide insights into true mutation effects. Our analysis provides a list of testable hypothesis, ordered according to the significance of mutation effect, that can be readily tested in laboratory to verify the real mutation effect in vitro. Besides, if a particular TF is known to regulate some gene and our analysis provides a big p-value change for the affinity of that TF due to mutation, this provides a strong evidence for the mutation effect and this should be taken into account when studying the disease

Table 1. Experimentally verified mutations and their effect on TF binding. p-values are presented only for those PSSMs that show relevant changes. wt=wildtype, Δp-value=(p-value of wt) $-$ (p-value of mutated sequence), mutation position is relative to TSS, MW=matrix width, POM= mutation position on matrix

gene symbol	mutation	mutation position	TF	MW	POM	effect on binding	Δp-value	p-value of wt	disease	reference
ALOX	A→G	-292	SPI1	6	2	increase	0.356	0.592	(anti)inflammatory effects	[4]
HBD	T→C	-77	GATA1	13	12	decrease	-0.386	0.553	δ-thalassemia	[3]
HBG2	C→G	-202	SP1	10	4	increase	0.274	0.540	hereditary persistence of fetal hemoglobin	[18]
HBG2	C→G	-202	SP1	10	5	increase	0.402	0.702	"	[18]
HBG2	C→G	-202	SP1	13	6	increase	0.658	0.861	"	[18]
HBG2	C→G	-202	SP1	10	4	increase	0.373	0.653	"	[18]
HBG2	C→G	-202	SP1	10	4	increase	0.206	0.420	"	[18]
PROC	T→C	-14	HNF-1	15	7	decrease	-0.216	0.265	protein C deficiency	[19]
UROS	C→A	-90	CP2	18	13	decrease	-0.207	0.143	congenital erythro-poietic porphyria	[20]
UROS	C→A	-90	CP2	11	11	decrease	-0.274	0.164	"	[20]
UROS	T→C	-70	GATA1	14	8	decrease	-0.317	0.085	"	[20]
UROS	T→C	-70	GATA1	13	7	decrease	-0.206	0.038	"	[20]

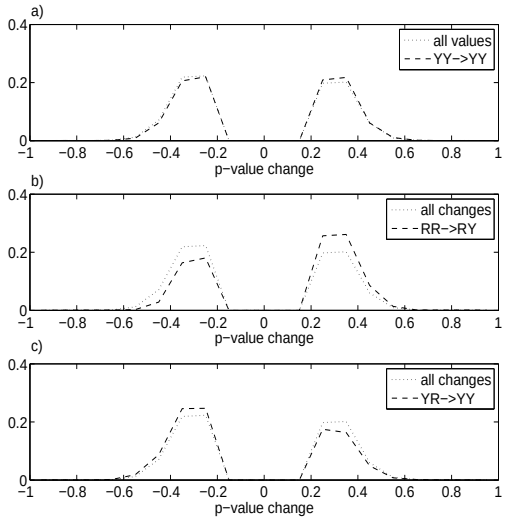

Figure 2. Distributions of the p-value changes in three different dinucleotide mutation types. a) YY→YY b) RR→RY c) YR→YR, Y=pyrimidine, R=purine

mechanisms on molecular level.

The distribution of the p-value changes of the relevant dataset can be seen in Figure 1b). It can be seen that the left side of the bimodal distribution has somewhat larger area than the right side i.e. the mutations cause more often the loss of TF binding affinity than create a new TFBS.

We computed the distributions of the p-value changes for each mutation type (16 dinucleotide classes). The distributions for different classes varied remarkably. In the Figure 2 is the distribution of the p-value changes in three different cases where mutation is in the second nucleotide. In all of the three plots there is also the distribution of the all p-value changes that exceeded the thresholds, as a reference distribution. It can be inferred based on the plots that the mutation type affects the binding affinity change differently. For mutations YY→YY (Figure 2a), the probability of formation a new TFBS is as probable as a disruption of an old binding site. This was also the case for mutation type RR→RR when mutation occurred in the second nucleotide and for RY→YY, RY→RY, YR→YR and YY→YY, if the mutation was in the first nucleotide. For mutations RR→RY and YY→YR, RR→YR and YY→RY the mutation more often caused a new binding site than disrupted an existing one (Figure 2b)). The rest of the mutations caused more likely the removal of an old binding site than making a new one as can be seen in an example in Figure 2c). The results suggest that purine-pyrimidine and pyrimidine-purine dinucleotides are in important roles in TF binding. It has been previously shown that pyrimidine-purine steps are flexible allowing the DNA strands to form sharp kinks [6]. This is important for TF which usually bends the DNA or binds to a bent DNA. Nevertheless, such flexibility is not shown to occur with all purine-pyrimidine steps. However, an RY step GC can also form more conformations than for example the AA and TT steps [7].

4. CONCLUSION

We have shown that regulatory mutations can change the TF binding affinity remarkably. This does not originate only from a single nucleotide mutation but also the type the surrounding nucleotides.

PSSMs are a widely used method to model TF binding. A big problem of PSSMs is, however, the number of false positives in predicting TFBSs. As our studies with experimentally verified TFBSs and the mutations affecting them showed, the PSSM modeling does not assign an extremely high p-values to TFBSs. This can be because of PSSM matrixes which does not have any corre-

lation between different bases. Our studies have shown that the dinucleotides in TFBSs affect the binding significantly. This is most likely caused by the ability of DNA strands to bend. Since different DNA-binding domains of TFs have different binding mechanisms and demands for DNA bending it could be more appropriate to study each TF family separately.

In the future it is important to incorporate additional knowledge into TF binding prediction. Previously, models that combine the nucleosome positions or Chromatin ImmunoPrecipitation on chip (ChIP-chip) data are shown to predict TF binding better than pure PSSMs [22], [23]. Other additional data sources can be also combined to models, for example DNase hypersensitive sites or conservation data. It should be also taken into account that in the cell, there is not just a single TF type present at a certain time, but the situation can be thought to be a competition between different TFs and other molecules to bind the DNA strand [21]. Thus, the TF binding differs in different states of the cell depending on the TFs present and their concentrations.

5. REFERENCES

[1] P. E. Newburger, D. G. Skalnik, P. J. Hopkins, A. A. Eklund, and J. T. Curnutte, "Mutations in the promoter region of the gene for gp91-phox in x-linked chronic granulomatous disease with decreased expression of cytochrome b558," *J Clin Invest*, vol. 94, pp. 1205–11, Feb 1994.

[2] J. Theuns, N. Brouwers, S. Engelborghs, K. Sleegers, V. B. V, E. Corsmit, T. de Pooter, C. M. van Duijn, P. P. de Deyn, and C. van Broeckhoven, "Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease," *Am J Hum Genet*, vol. 26, pp. 936–46, Jun 2006.

[3] M. Matsuda, N. Sakamoto, and Y. Fukunaki, "δ-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the δ-globin gene promoter," *Blood*, vol. 80, pp. 1347–51, 1992.

[4] J. Wittwer, J. Marti-Juan, and M. Hersberg, "Functional polymorphism in ALOX15 results in increased allele-spesific transcription in macrophages through binding of the transcription factor SPI1," *Hum Mutat*, vol. 27, no. 2, pp. 78–87, 2006.

[5] M. C. Andersen, P. G. Engström, S. Lithwick, D. Arenillas, P. Eriksson, B. Lenhard, W. W. Wasserman, and J. Odeberg, "In silico detection of sequence variations modifying transcriptional regulation," *PLoS Comput Biol*, vol. 4, pp. e5, Jan 2008.

[6] M. Suzuki, D. Loakes, and N. Yagi, "DNA conformation and its changes upon binding transcription factors," *Adv Biophys*, vol. 32, pp. 53–72, 1996.

[7] A. A. Travers, "The structural basis of DNA flexibity," *Philos Transact A Math Phys Eng Sci*, vol. 15, pp. 1423–38, Jul 2004.

[8] A. Sarai and H. Kono, "Protein-DNA recognition patterns and predictions," *Annu Rev Biophys Biomol Struct*, vol. 34, pp. 379–98, 2005.

[9] R. E. Harrington, "DNA curving and bending in protein-dna recognition," *Mol Microbiol*, vol. 6, pp. 2549–55, Sep 1992.

[10] C. O. Pabo and R. T. Sauer, "Transcription factors: Structural families and principles of DNA recognition," *Annu Rev Biochem*, vol. 61, pp. 1053–95, 1992.

[11] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. Thomas, S. Abeysinghe, M. Krawczak, and D. N. Cooper, "The Human Gene Mutation Database (HGMD ®): 2003 update," *Hum Mutat*, vol. 21, pp. 577–581, 2003.

[12] G. D. Stormo, "DNA binding sites:representation and discovery," *Bioinformatics*, vol. 16, pp. 1416–23, Jan 2000.

[13] R. Staden, "Computer methods to locate signals in nucleic acid sequences," *Nucleic Acids Res*, vol. 12, pp. 505–519, Jan 1984.

[14] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, P. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374–8, 2003.

[15] A. Sandelin, Walkema, P. Engström, W. Wasserman, and B. Lenhard, "JASPAR: an open access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Res*, vol. 32, pp. D95–7, 2004.

[16] B. Lenhard and W. Wasserman, "TFBS: Computational framework for transcription factor binding site analysis," *Bioinformatics*, vol. 18, pp. 1135–6, 2002.

[17] J. V. Ponomarenko, G. V. Orlova, M. P. Ponomarenko, S. V. Lavryushev, and T. I. Merkulova, "rSNP_Guide: a database documenting influence of substitutions in regulatory gene regions onto their interaction with nuclear proteins and predicting protein binding sites, damaged or appeared de novo due to these substitutions," in *Proceedings of BGRS'2000*, 2000, pp. 69–72.

[18] F. S. Collins, C. J. jr Stoeckert, G. R. Serjeant, B. G. Forger, and S. M. Weissman, "G gamma beta+ hereditary persistence of fetal hemoglobin: cosmid cloning and identification of a specific mutation 5' to the G gamma gene," *Proc Natl Acad Sci U S A*, vol. 81, pp. 4898–8, Aug 1984.

[19] L. P. Berg, D. A. Scopes, A. Alhaq, V. V. Kakkar, and D. N. Cooper, "Disruption of a binding site for hepatocyte nuclear factor 1 in the protein C gene promoter is associated with hereditary thrombophilia," *Hum Mol Genet*, vol. 3, pp. 2147–52, Dec 1994.

[20] C. Solis, G. I. Aizencan, K. H. Astrin, D. F. Bishop, and R. J. Desnick, "Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria," *J Clin Invest*, vol. 107, pp. 753–62, Mar 2001.

[21] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, "Predicting expression patterns from regulatory sequence in Drosophila segmentation," *Nature*, vol. 451, pp. 535–540, Jan 2008.

[22] L. Narlikar, Raluca, and A. J. Hartemink, "A nucleosome-guided map of transcription factor binding sites in yeast," *PLoS Comput Biol*, pp. 2199–208, 2007.

[23] H. Kim, K. J. Kechris, and L. Hunter, "Mining discriminative distance context of transcription factor binding sites on ChIP enriched regions," in *ISBRA*, 2007, pp. 338–49.