Knowledge-Based Systems 15 (2002) 225–233

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

Mining DNA sequences to predict sites which mutations cause genetic diseases

Julia Ponomarenko, Tatyana Merkulova, Galina Orlova, Oleg Fokin, Elena Gorshkov, Mikhail Ponomarenko*

Institute of Cytology and Genetics, 10 Lavrentyev Ave., 630090 Novosibirsk, Russia

Received 15 March 2001; revised 2 May 2001; accepted 31 May 2001

Abstract

Currently single nucleotide polymorphism (SNP) analysis becomes the crossroad of bioinformatics and medicine. We have developed a data mining system, http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/, called rSNP_Guide, to discover regulatory sites in DNA sequences, which mutations could be the cause of genetic diseases. During the first step, we estimate the abilities of the proteins considered to bind to genomic DNA, which alterations by mutations are associated with a genetic disease under study. During the second step, we formalize the disease-associated experimental data on the SNP-referred alterations in DNA binding to unknown protein. During the third step, we cluster fuzzily all known proteins examined so that to determine one of them, which specific site is altered by mutations in consistence with that of the unknown protein experimentally associated with genetic disease. During the fourth step, we predict the known protein, which binding site is (i) resent on DNA and (ii) altered by mutations associated with genetic disease. Finally, during the last step, we estimate the robustness of this prediction. The rSNP_Guide has been tested on the SNPs with the known relationships between regulatory site alterations and genetic disease penetration. Besides, the novel SNPs-referred regulatory sites associated with the genetic disease penetrations were discovered and, then, successfully confirmed experimentally. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Data mining; Single nucleotide polymorphism; Regulatory site; Mutation; Genetic disease

1. Introduction

Since human genome is sequenced in draft, single nucleotide polymorphism (SNP) analysis became the crossroad of medicine and bioinformatics: over 2.84 millions of SNPs are detected, including 1.42 millions of SNPs mapped [1]. Thus, it is necessary to develop the systems predicting the regulatory sites, SNP-referred alterations in that are associated with genetic diseases. For recognizing the natural sites on the basis of their textual similarities, a number of the pattern recognition tools has been developed earlier [2]. Nevertheless, some experiments [3] have demonstrated that the sites damaged by mutations could not be reliably recognized only by their similarity to the known sites, because defective sites, as a rule, have no natural analogs.

On the other hand, due to the drastic growth in numbers of genome databases and on-line publications, the novel data-mining tools were developed for automated extracting and accumulating many site-referred knowledges [4]. By keeping this in mind, heterogeneity of compilations of site sequences becomes visible, and, hence, many novel tools clustering the site-related data into homogeneous subsets were reported [5]. Besides, novel knowledge discovery tools has been developed for revealing unobvious regularities in the structure of site neighborhoods in regulatory gene regions. These regularities were earlier ignored, thus, limiting the site recognition accuracy [6]. Finally, by integrating pattern recognition and data mining tools, “in silico” biology systems were developed for in-depth studying of genome structure, function, and variation data [7]. Thus, during the so-called “Post Genome Era”, bioinformatics meets with data mining [8].

Following this way, we have used the generating hypotheses [9], decision making [10], and fuzzy sets [11] for studying the textual [12], physicochemical and conformational [13] regularities of the true site sequence-activity relationships [14] and positionings on DNA [15]. By cross-validation testing, we have shown that both regressional and positional regularities of a given site are comparable [16].

In this paper, we present our data mining system, http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/, called rSNP_Guide, predicting regulatory sites in DNA sequences, which mutations could be responsible for genetic disease

[*] Corresponding author. Tel.: +7-3832333119; fax: +7-3832331278. *E-mail address:* jpon@bionet.nsc.ru (M. Ponomarenko).

PII: S0950-7051(01)00144-7

Table 1
Relationships between genetic disease and the SNP-referred alterations in DNA sequence and binding to unknown proteins (mutation, **bold-faced and underlined**)

Genetic disease	Gene, region	Mutations	SNP		Genome DNA sequence	DNA/protein binding
Tumors in lung [17]	K-*ras*, intron #2	Wild Type	CA	(+)	5′-gaaaCtccacttAtca-3′	Present
				(−)	5′-tgaTaagtggaGtttc-3′	
		296A/*C*	CC	(+)	5′-gaaaCtccactt**C**tca-3′	Absent
				(−)	5′-tga**G**aagtggaGtttc-3′	
		288C/*G*, 296A/*C*	GC	(+)	5′-gaaa**G**tccactt**C**tca-3′	Absent
				(−)	5′-tga**G**aagtgga**C**tttc-3′	
Drug dependence, Tourette syndrome, attention deficit [18]	*TDO2*, intron #6	Wild Type	WT	(+)	5′-ataatgGcaGataaga-3′	Present
				(−)	5′-tcttatCtgCcattat-3′	
		663G/*A*	M1	(+)	5′-ataatg**A**caGataaga-3′	Absent
				(−)	5′-tcttatCtg**T**cattat-3′	
		666G/*T*	M2	(+)	5′-ataatgGca**T**ataaga-3′	Reduced
				(−)	5′-tcttat**A**tgCcattat-3′	
Severe malaria [19]	*NTFα*, promoter	Wild Type	αG	(+)	5′-tgtctggaaGttagaa-3′	Absent
				(−)	5′-ttctaaCttccagaca-3′	
		-376G/*A*	αA	(+)	5′-tgtctggaa**A**ttagaa-3′	Present
				(−)	5′-ttctaa**T**ttccagaca-3′	
Type I protein C deficiency [20]	*pC*, promoter	Wild Type	WT	(+)	5′-ggttatggaCtaactc-3′	Present
				(−)	5′-gagttaGtccataacc-3′	
		-114C/*T*	MT	(+)	5′-ggttatgga**T**taactc-3′	Absent
				(−)	5′-gagtta**A**tccataacc-3′	
Bernard-Soulier syndrome [21]	*GpIbβ*, promoter	Wild Type	WT	(+)	5′-gtgctatCtgccgctg-3′	Present
				(−)	5′-cagcggcaGatagcac-3′	
		-133C/*G*	MT	(+)	5′-gtgctat**G**tgccgctg-3′	Absent
				(−)	5′-cagcggca**C**atagcac-3′	
Severe bleeding disorder [22]	*fVII*, promoter	Wild Type	WT	(+)	5′-cccctccCccatccct-3′	Present
				(−)	5′-agggatggGggagggg-3′	
		-94C/*G*	MT	(+)	5′-cccctcc**G**ccatccct-3′	Absent
				(−)	5′-agggatgg**C**ggagggg-3′	
Hereditary persistence of fetal hemoglobin [23]	*Gγ*, promoter	Wild Type	WT	(+)	5′-ccttgacCaatagcct-3′	Present
				(−)	5′-aggctattGgtcaagg-3′	
		-114C/*T*	MT	(+)	5′-ccttgac**T**aatagcct-3′	Absent
				(−)	5′-aggctatt**G**gtcaagg-3′	

susceptibility/resistance. Our system adopts the following approach.

1. During the first step, we estimate the abilities of the proteins considered to bind to genomic DNA, which alterations by mutations are associated with a genetic disease under study.
2. During the second step, we formalize the disease-associated experimental data on the SNP-referred alterations in DNA binding to unknown protein.
3. During the third step, we cluster fuzzily all known proteins examined so that to determine one of them, which specific site is altered by mutations in consistence with that of the unknown protein experimentally associated with genetic disease.
4. During the fourth step, we predict the known protein, which binding site is (i) resent on DNA and (ii) altered by mutations associated with genetic disease.
5. Finally, during the last step, we estimate the robustness of this prediction. The rSNP_Guide has been tested on the SNPs with the known relationships between regulatory site alterations and genetic disease penetration.

This paper is divided into the following sections. First, we introduce a biological background with proper mathematical definitions. Next, by using these definitions, we discuss implementation of the data mining algorithm in the rSNP_Guide system and its design. In what follows, we give the experimental results of the rSNP_Guide application to studying of the K-*ras* gene (associated with tumor in lung) [17] and *TDO2* gene (associated with mental disorders) [18], which mutations alter DNA sites binding with unknown proteins. Also, rSNP_Guide was implemented for analysis of some other genes, *NTFα* [19], *pC* [20], *GpIbβ* [21], *fVII* [22], and *Gγ* [23] with the known site alterations causing genetic diseases. Finally, we discuss the rSNP_Guide future development.

2. Biological background

Since replication, translation and other vitally important

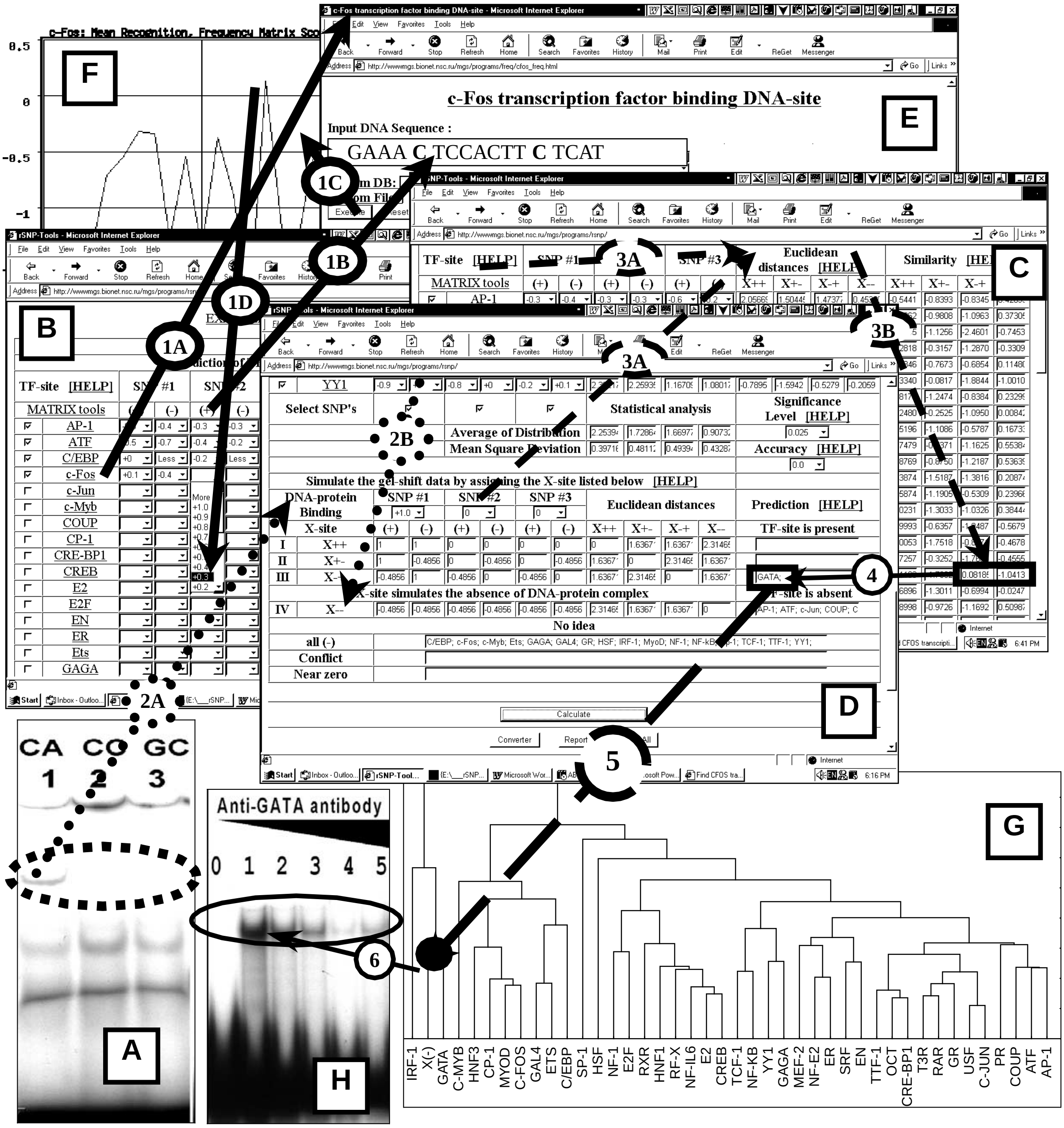
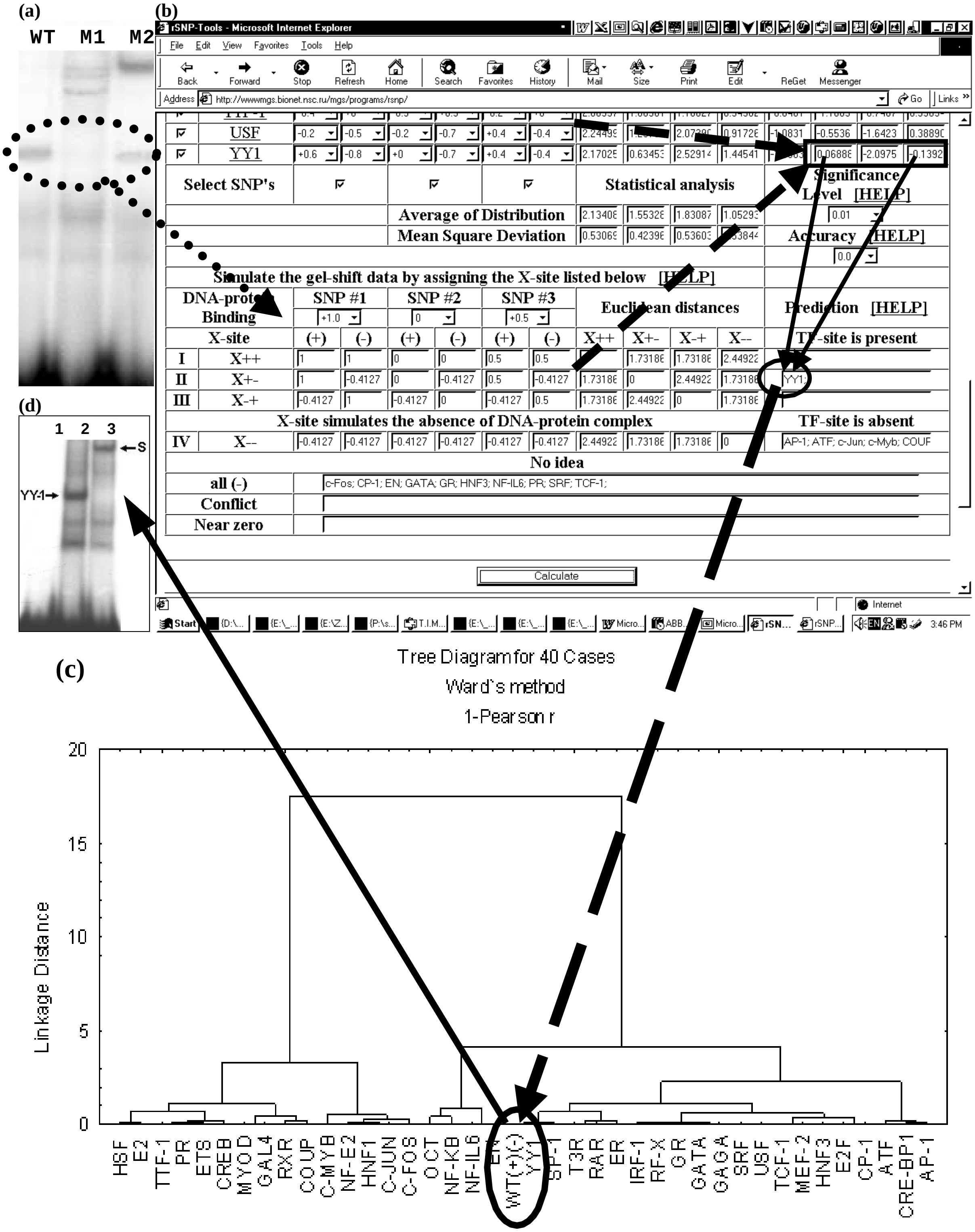


Fig. 1. Data mining approach: (i) Estimating DNA/protein-binding abilities (arrows 1A, 1B, 1C, and 1D); (2) Formalizing the DNA/protein-binding pattern (arrows 2A and 2B); (iii) Fuzzy clustering the proteins (arrows 3A and 3B); (iv) Predicting (arrow 4); (v) Testing of robustness (arrow 5). Arrow 6, the control experiment test was planned on the basis of the TF-site predicted by the rSNP_Guide. Screens: A, the SNP-referred experimental data on the alterations in the examined DNA binding to unkown protein (input data); B, C, and D, the rSNP_Guide user's interface; E, the user interface of a TF-site recognition tools loaded by rSNP_Guide; F, the tools output; G, the output window of the standard package STATISTICA, preference of which as a the robust test tools platform could be addressed to its common acceptance; H, the control experiment, anti-GATA antibody of which proved the GATA-site prediction. Lines: 0—no extract; lines 1, 2, 3, 4, and 5—lung extract pre-incubated with the gradient concentrations of the anti-GATA antibodies. As one can see, the band under study is decreasing with the increase of the anti-GATA antibody concentration. Thus, the site GATA present in the allele "CA" is associated with the susceptibility to tumors in lung, such as it has been predicted by rSNP_Guide.

molecular genetic processes are controlled by the so-called "regulatory proteins" binding to specific genome DNA sites, the SNP-referred mutations of DNA produce alterations in DNA binding to regulatory proteins, thus, causing genetic diseases.

The system rSNP_Guide presented analyzes two sorts of experimental data: (i) DNA sequences and (ii) alterations in DNA-protein binding pattern [24]. These data sets are exemplified in Table 1 and in Figs. 1 and 2. As seen in Table 1, each SNP-referred variant is represented by two DNA sequences corresponding to two oppositely directed strands of the double helix DNA: the attribute "(+)-chain"



☑	USF	-0.2	-0.5	-0.2	-0.7	+0.4	-0.4	2.2449	1.2	2.073	0.9172	-1.0831	-0.5536	-1.6423	0.3889
☑	YY1	+0.6	-0.8	+0	-0.7	+0.4	-0.4	2.17025	0.6345	2.5291	1.44541	[illegible]	0.0688	-2.0975	-0.1392
Select SNP's		☑		☑		☑		Statistical analysis				Significance Level [HELP]			
				Average of Distribution				2.1340	1.5532	1.83087	1.0529	0.01			
				Mean Square Deviation				0.5306	0.4239	0.5360	3844	Accuracy [HELP]			
												0.0			
Simulate the gel-shift data by assigning the X-site listed below [HELP]															
DNA-protein Binding		SNP #1 +1.0		SNP #2 0		SNP #3 +0.5		Euclidean distances				Prediction [HELP]			
X-site		(+)	(-)	(+)	(-)	(+)	(-)	X++	X+-	X-+	X--	TF-site is present			
I	X++	1	1	0	0	0.5	0.5		1.7318	1.7318	2.4492				
II	X+-	1	-0.4127	0	-0.4127	0.5	-0.4127	1.7318	0	2.4492	1.7318	YY1;			
III	X-+	-0.4127	1	-0.4127	0	-0.4127	0.5	1.7318	2.4492	0	1.7318				
X-site simulates the absence of DNA-protein complex												TF-site is absent			
IV	X--	-0.4127	-0.4127	-0.4127	-0.4127	-0.4127	-0.4127	2.4492	1.7318	1.7318	0	AP-1; ATF; c-Jun; c-Myb; COUF			
No idea															
all (-)		c-Fos; CP-1; EN; GATA; GR; HNF3; NF-IL6; PR; SRF; TCF-1;													
Conflict															
Near zero															

Calculate

Fig. 2. Binding of the nuclear proteins to the oligonucleotides corresponding to the alleles WT, M1 and M2 of #6 intron *TDO2* gene: (a) the SNP-referred experimental data on the alterations in the examined DNA binding to unknown protein (the main band and its analysis steps are marked by arrows); (b) how to use rSNP_Guide in the case of these data analysis, both algorithm and arrow definition of which are the same as the caption to Fig. 1; (c) an example of the robust test carried out; (d) the control test by anti-YY1 antibody super-shift assay: 1—no extract; 2—liver nuclear extract, no antibodies; 3—liver nuclear extract was pre-incubated with anti-YY1 antibodies. As seen, the main band, association of which with YY1-site has been predicted by the rSNP_Guide (arrow YY1), is super-shifted by the pre-incubation with anti-YY1 antibodies (arrow S).

denotes DNA strand, which serves as a matrix for translation and protein synthesis, whereas the attribute "(−)-chain" marks complementary mirror sequence located at the opposite strand of DNA helix.

Fig. 1(A) exemplifies another type of the SNP-referred experimental data analyzed by SNP_Guide. It is compiled of alterations in DNA binding pattern to unknown protein. These experimental data are obtained as follows: under definite experimental conditions, short oligoDNAs corresponding to genome variants under study are synthesized and preincubated together with the mixture of all proteins extracted from cell nucleus. By electrophoresis, the complexes of these oligoDNAs with unknown proteins are separated according to their mobility in the gel (Fig. 1(A)). As one can see in Fig. 1(A), three SNP-referred genome variants are represented by three vertical lines in four horizontal band series, which characterize four patterns of DNA binding to unknown proteins. These patterns are differing by mobility of DNA-protein complexes in the gel. In this figure, the dotted frame marks the SNP-referred alterations in the examined pattern of DNA binding to unknown protein: the band characterizing genome variant "CA" absents in both "CC" and "GC" variants. The rSNP_Guide analyzes these data under presumption that the SNP-referred alterations in the examined pattern of DNA binding to unknown protein refer to a single target protein. This target protein strongly binds to DNA due to the presence of its specific site, which was altered by mutations studied. However, other proteins extracted from a cell nucleus could only weakly bind to DNA examined, because their specific binding sites are absent there. Hence, all regulatory proteins extracted from cell nucleus should be similar by their DNA-binding estimates, which, in turn, strictly differ from the estimate for the particular target protein. Thus, this target protein could be detected by means of clusterization of the proteins to each other.

3. Mathematical definitions

To describe the data mining approach implementing the above biological background by rSNP_Guide system, the following mathematical definitions were introduced.

TF, variable attribute, the abbreviation of "Transcription Factor", which is the type of regulatory protein.

N, constant value, the number of SNPs under study.

$\mathrm{Score}_n(\mathrm{TF})$, variable value, the DNA/protein-binding rate estimated upon the n-th DNA sequence variant by using TF-protein specified procedure, the resulted values of which are normalized by the rule: (i) the means averaged upon 1000 random DNAs equal to "$-$ 1"; (ii) the means averaged over all known true TF-site sequences equal to "$+$ 1", (iii) the threshold discriminating between the true TF-sites and random DNA sequences equals to "0".

$X - - = \{x_{--n}\}_{1\le n\le 2N}$, variable vector, the cluster center aimed at collecting all nuclear proteins except the only target protein, which specific site on DNA is altered by mutations under study.

$X + + = \{x_{++n}\}_{1\le n\le 2N}$, variable vector, the clustering center aimed at detecting the single target protein in case if "both DNA strands carry the sites altered by mutation and specific for binding of this target protein".

$X + - = \{x_{+-n}\}_{1\le n\le 2N}$, variable vector, the clustering center aimed at detection of the single target protein in case if "only the (+)-chain has the mutation-altered sites specific for binding to this target protein".

$X - + = \{x_{-+n}\}_{1\le n\le 2N}$, variable vector, the clustering center aimed at detection of single target protein in case if "only the (−)-chain carries the sites altered by mutation and specific for binding to this target protein".

$X + \cdot \in \{X + +, X + -, X - +\}$, all three possible clustering centers aimed at detection of the target protein; set of which is formalized by SNP-referred alterations in the pattern of DNA binding to unknown protein. It is measured as the relative degree on the scale from "$-$ 1" (minimal) to "$+$ 1" (maximal) with the "0" in the case "no ideas".

$D_{\mathrm{TF},X}$, variable value, Euclidean distance from each clustering center, Xs, to each vector representing DNA binding to known protein, TF, dependent on the SNP-referred sequence variants.

$D_{X--,X+\cdot}$, variable values, the membership thresholds discriminating the clustering center aimed at detection of the target protein, $X + \cdot$, from the clustering center, $X--$, aimed at collecting the rest proteins.

$d_{\mathrm{TF},X+\cdot}$, variable value, a clustering memberships estimate, which states that "on DNA examined, there exists the site altered by mutation and specific for binding with the protein TF".

$d_{\mathrm{TF},X--}$, variable value, a cluster memberships estimate, which states that "on DNA examined, there exists no sites altered by mutation and specific for binding with the protein TF".

$t_{\alpha;\nu}$, constant value, the Student's t-coefficient at significance level α and freedom degree ν.

s.d.(ξ), variable value, an estimate of the standard deviation of the ξ-variable.

$\{X + \cdot \approx \mathrm{TF}\}$, variable value, prediction of the type "TF-site, alteration of which is associated with genetic disease".

4. First step of data mining: estimating DNA-protein binding ability

With 2 × N DNA sequences of N SNP-referred variants of interest for the (+) and (−) strands prepared in advance, a user loads the rSNP_Guide, http://wwwmgs.bionet.nsc.ru/mgs/programs/rsnp/. Screen B in Fig. 1 shows the upper section of the user interface window, where a protein (TF) of interest should be selected. When a TF is clicked on (arrow 1A), the tools appear for its corresponding the TF-specific site recognition (Screen E). The user enters into the input box each sequence variant (arrow 1B) and receive the graphical representation of the TF site recognition Score (Screen F, arrow 1C). With a dominant peak, the corresponding score on the left axis is entered in the proper box in the user interface (arrow 1D). When all sequence variants have been treated, the next TF is selected and the process repeated. When all TFs of interest have been examined by this way, the vectors, $\{\mathrm{Score}_n(\mathrm{TF})\}_{1\le n\le 2N}$, are assigned by the numerical estimates of the SNP-referred DNA binding to each examined protein, TF, and, thus, this step is ended.

5. Second step of data mining: formalizing DNA/protein-binding pattern

When all TFs of interest have been examined, the SNP-referred experimental data on the alterations in the examined DNA binding to unknown protein are entered in the interface section "DNA/protein Binding" (Fig. 1, Screen D) data for each sequence variant, estimating the relative degree of protein binding on a scale of $+1$ (maximal) to -1 (minimal). Screen A in Fig. 1 exemplifies the experimental data. Formalizing of these data is illustrated in Screen D: since the framed band corresponding to DNA variant "CA" is absent in genome variants "CC" and "GC", the box "SNP #1" is assigned by the maximal relative degree " + 1", whereas the boxes "SNP #2" and "SNP #3" are assigned by the value "0". Next, these data are automatically input into the proper boxes of three vectors $X{+}{+}$, $X+\ -$, and $X-\ +$, which are located just below the interface section "DNA/protein Binding". After that, the means averaged upon all negative estimates of the examined DNA binding to TFs of interest are automatically calculated and put into the remaining boxes of three vectors $X{+}{+}$, $X+\ -$, and $X-\ +$, and also, into each box of the vector $X-$ located below. As one can see, when all components of four Xs are assigned, four possible cases of "the examined DNA binding to unknown protein due to the presence/absence of this protein-specific site on $(+)$ and/or $(-)$ DNA strands are formalized, and, thus, the second step of data mining is fulfilled.

6. Third step of data mining: fuzzy clustering regulatory proteins

At this step, we compare two sorts of DNA-protein binding estimates prepared at two preceding steps: (i) if the known proteins are examined, $\{\mathrm{Score}_n(\mathrm{TF})\}_{1\le n\le 2N}$, and, (ii) if unknown target protein binds to the DNA site altered by mutations, Xs. First, Euclidean distances, $D_{\mathrm{TF},X}$, scaling a similarity between each pair (TF; X) is calculated:

$$D_{\mathrm{TF},X} = [\Sigma_{1\le n\le 2N}(\mathrm{Score}_n(\mathrm{TF}) - x_n)^2]^{1/2}. \tag{1}$$

In Fig. 1, arrow 3A illustrates the automated calculation by the formula (1). The results of this calculation could be checked up by a user. Next, three thresholds, $D_{X--,X+\cdot}$, discriminating each target clustering nuclei, $X+\cdot \in \{X++,\ X+-,\ X-+\}$, from the main clustering nucleus $X-$ are calculated by formula (1).

Finally, on the basis of these distances and their thresholds, for each known protein examined, TF, for its membership rate estimates, $d_{\mathrm{TF},X}$, of each cluster representing one of four possible cases "the TF-specific site is present/absent on the $(+)/(-)$-chain of the DNA altered by the mutations studied" are calculated by using t-test, as it is shown by arrow 3B in Fig. 1:

$$d_{\mathrm{TF},X--} = t_{\alpha;\nu} \times \mathrm{s.d.}(D_{\mathrm{TF},X--}) - D_{\mathrm{TF},X--}; \tag{2}$$

$$d_{\mathrm{TF},X+\cdot} = D_{X--,X+\cdot} - D_{\mathrm{TF},X+\cdot} - t_{\alpha;\nu} \times \mathrm{s.d.}(D_{\mathrm{TF},X+}). \tag{3}$$

As seen, since each known regulatory protein, TF, has been assigned by its membership rate estimate to each cluster center formalizing the presense/absence on DNA the TF-specific site, which was altered by mutations under study, this step of data mining is fulfilled in accordance with fuzzy sets criterion [11].

7. Fourth step of data mining: predicting site associated with disease

During this data mining step, only the single known regulatory protein, TF, would be detected by one of three vectors, $X+\cdot \in \{X++,\ X+-,\ X-+\}$, formalizing "the alterations in DNA site-specific binding to unknown protein", when all remaining proteins could be clustered to each other by the remaining vector, $X-$, formalizing the case of "no specific site is present on this DNA altered by mutations under study". With all membership rate estimates available, for each pair $\{X+\cdot, \mathrm{TF}\}$, use the rule:

$$\mathrm{IF}\ \ \{d_{\mathrm{TF},X+\cdot} > 0\}\ \ \mathrm{AND}\ \ \{d_{\mathrm{TF},X--} < 0\}\ \ \mathrm{THEN}\ \ \{X+\cdot \approx \mathrm{TF}\} \tag{4}$$

In Fig. 1, arrow 4 illustrates successful usage of the rule in the real case of the GATA site presence in the intron #2 of the mouse K-*ras* gene, which is associated with the lung tumor.

If either several or no TFs are predicted, the significance value ($\alpha < 0{,}025$, default) can be varied in between 0.00005 and 0.1. In addition, the handmade threshold of the cluster membership rate estimates could be varied from 0 (default) to 1 in accordance with degree of similarity between the known proteins and the unknown one, which was detected due to alterations of its site by mutation.

8. Fifth step of data mining: testing robustness

For the testing of robustness [25], the vectors $\{\mathrm{Score}_n(\mathrm{TF})\}_{1\le n\le 2N}$ of the DNA binding estimates of all proteins considered (Step 1) and only one cluster center,

$X + \cdot$, selected by the rule (4) are inputted into the standard package STATISTICA. With these input data, each STATISTICA option-pair, one among six similarity scales and one among seven clustering methods, is tested for clarifying whether preliminary prediction $\{X + \cdot \approx \mathrm{TF}\}$ is confirmed (i.e. the test is robust) or not. When all $7 \times 7 = 49$ "scale + method" pairs are checked up, the ratio robust/all tests estimates the robustness of the $\{X + \cdot \approx \mathrm{TF}\}$ prediction.

9. System design

Following the modern concept "free source code", our data mining approach is implemented with the system rSNP_Guide by the standard Java-script tools, the source code of which is loaded and, after that, both line-by-line interpreted and executed by the Web-browser of a user's platform. We have successfully tested that these simple Java-script tools are compatible with any Web-browser available for us. The rSNP_Guide is the so-called "real-time" system, i.e. a single user's query is immediately processed. Depending on genome sequence variants and regulatory proteins of a user interest, the in-depth analysis of given SNP-referred experimental data by the rSNP_Guide requires from several minutes to many hours. Finally, for maximizing a user's trust to predictions made by rSNP_Guide, all results of computations are available (as seen in Figs. 1 and 2), and, thus, both intermediate and final results could be verified by user with a simplest calculator used "by-hand".

10. Experimental results

K-*ras* gene. The K-*ras* gene is widely used as genetic marker of susceptibility in different mouse strains to spontaneous and chemically induced mutagenesis in lung[17]. Three alleles of this gene are known. They are denoted as sensible (K^s), intermediate (K^i), and resistant (K^r) alleles and are related to different expression patterns of this gene. All K^r allele carriers are characterized by tandem repeat of 37 bp in length in the second intron (282–355 bp). K^s and K^i carriers have only a single copy of this repeat. In addition, we have found two single nucleotide substitutions inside the repeated unit, which correlate to lung tumor susceptibility. In particular, the tumor-susceptible allele has the C nucleotide at 288 bp position and A nucleotide at 296 bp position, whereas the intermediately resistant allele is characterized by substitution ? ← C, and the resistant allele carries one more substitution ? ← G [17]. We have supposed that nucleotide substitutions may be located within the region binding to some regulatory protein, thus leading to decrease in the K-*ras* gene expression.

For examining this supposition, in this work we have synthesized three 30 bp double stranded oligonucleotides corresponding to three alleles within the region between 278–307 bp of the second intron of the mouse K-*ras* gene: the tumor-sensible "CA", 5′-gtgcaagaaa**C**tccactt**A**tcatgagagct-3′; the tumor-intermediate "CC", 5′-gtgcaagaaa**C**tccactt**C**tcatgagagct-3′; and, finally, the tumor-resistant "GC", 5′-gtgcaagaaa**G**tccactt**C**tcatgagagct-3. Then, under fixed concentration, each oligonucleotide has been preincubated together with the lung nuclear extract prepared by the standard manner that is the mixture of all regulatory proteins presenting within the nucleus of the lung cells. For identifying the complexes formatted due to the olignucleotide binding to the lung nuclear proteins, the separation of these complexes by their molecular mass, which influences gel mobility, was made by electrophoresis. The result is given in Fig. 1(A). In this figure, one can see four horizontal band series demonstrating that at least four regulatory proteins from lung cell nucleus can bind to this particular DNA region. In addition, the dotted frame in this figure demonstrates that the main band characterizing the allele "CA" is absent in both "CC" and "GC" alleles, hereby three remaining band series remain constant within each allele. It means that the DNA region under study has four regulatory protein binding sites as minimum, whereas the only one of them is present in the tumor-sensible allele and absent in both intermediate and resistance alleles.

With these input data, rSNP_Guide has predicted the GATA site among 41 known sites considered (Fig. 1). To verify this GATA site prediction, we have planned and carried out one more gel mobility shift experiment, an additional step of which with respect to the first one was in incubating the lung nuclear extract together with the anti-GATA antibodies. These control test results are shown in Fig. 1(H). In this figure, one can see that the band corresponding to the tumor-sensible allele becomes less intensive with increase of the anti-GATA antibody concentration. It means that this band corresponds to the complex of the oligonucleotide "CA" binding to the regulatory proteins GATA, concentration of which in the lung nuclear extract decreased due to addition of the anti-GATA antibody. Thus, the gel mobility immune-shift assay has proved that the lung tumor susceptibility is dependent on the GATA site presence in the "CA" allele of the K-*ras* gene. So, resistance to lung tumor is associated with the absence of the GATA site.

TDO2 gene. One more example of our original SNP-related experimental research is application of the rSNP_Guide to the SNP-analysis for the *TDO2* gene, polymorphism of which has been identified and significantly positively associated with the drug dependence, Tourette syndrome, and attention deficit hyperactivity disorder [18]. For the *TDO2* gene, Fig. 2 illustrates (a) the SNP-referred experimental data on the alterations in the examined DNA binding to unknown protein; (b) both SNP-referred data input and predicted YY1-site output; (c) the robust test of this YY1 site prediction; and (d) the anti-YY1 antibody super-shift assay that completely

Table 2
Control tests results obtained by the system rSNP_Guide on the SNPs-related experimental data given in Table 1

Data			Prediction					Control test	
Gene	SNP	Extract/cell	N + [a]	Site	Chain	$d_{\mathrm{TF},X}$	N_R [b]	Experiment	Genetic disease
K-*ras*	CA	Lung	0.16	*GATA*	(+)	0.01	0.83	Anti-GATA antibody (this work)	Tumors in lung [17]
	CC								
	GC								
TDO2	WT	Liver	0.24	YY1	(−)	0.08	0.83	Anti-YY1 antibody (this work)	Drug dependence, Tourette syndrome attention deficit [18]
	M1								
	M2								
NTFα	αG	MonoMac6	0.10	OCT	(+)	0.07	0.6	Anti-OCT antibody [19]	Severe malaria [19]
	αA								
Pc	WT	HepG2	0.16	HNF1	Both	0.28	1.00	Anti-HNF1 antibody [20]	Type I protein C deficiency [20]
	MT								
GpIbβ	WT	CHRF-288	0.27	GATA	(−)	0.02	1.00	Anti-GATA antibody [21]	Bernard-Soulier syndrome [21]
	MT								
fVII	WT	HepG2	0.35	Sp1	(−)	0.01	0.71	Anti Sp1 antibody [22]	Severe bleeding disorder [22]
	MT								
Gγ	WT	MEL	0.35	CP-1	(+)	0.03	0.81	Consensus competition [23]	Hereditary persistence of fetal hemoglobin [23]
	MT								

[a] N + , ratio “false positives”/total recognitions of the examined regulatory sites upon all DNA sequence variants considered by these site textual patterns BEFORE the usage of the clustering approach

[b] N_R, ratio robust/total tests’ relying to the all possible “scale-method” pairs of six similarity scales and seven clustering methods available within the standard package STATISTICA.

supports the computer-assisted prediction by rSNP_Guide. As one can see in Fig. 2(d), for the *TDO2* gene, we have used another sort of antibodies than in case of the K-*ras* gene analysis (Fig. 1(H)). Addition of the anti-YY1 antibodies does not modify the binding pattern of DNA to the regulatory protein YY1. However, the molecular mass and gel mobility of the YY1/DNA-complex are altered because of addition of the anti-YY1 antibodies to the protein-oligonucleotides complex. As one can see in Fig. 2(d), the band under study is seen only in the liver nuclear cell extract (marked by arrow “YY1”) and it is shifted due to the anti-YY1 antibody additives (arrow “S”). This widely used method called “immune-supershift” has precisely proved that this band corresponds to the site-specific YY1/DNA complex formation. Thus, we have successfully predicted that mutations damaging the YY1 site are responsible for several mental disorders [18].

Genes with known site-genetic disease relationships. For control testing of the rSNP_Guide work, we have additionally studied several genes: *NTFα* [19], *pC[20]*, *GpIbβ* [21], *fVII* [22], and *Gγ* [23] with known site-genetic disease relationships (Table 1). The results obtained are presented in Table 2. As follows from the table, all control results are in agreement with the earlier published control test results obtained by using either antibody super-shift assay or competitor-binding experiments, as well as with almost all cluster-analysis schemes produced by the standard STATISTICA package and used for the testing of robustness.

11. Conclusion and perspectives

In this paper we have presented a data mining system, called rSNP_Guide, to discover regulatory sites in DNA sequences which mutations could be the cause of genetic diseases. Our Systems follows the following steps: (i) estimating the examined DNA binding to the known proteins of interest; (ii) formalizing the experimental data on the alterations in the DNA binding to unknown protein; (iii) fuzzy clustering the known proteins to each others in order to detect only one of them, which specific site is altered by mutations and consistent with unknown protein associated with genetic disease; (iv) predict the known protein, which specific site can be associated with genetic disease; (v) estimating the robustness of this prediction.

Our system is implemented in Java-skript and integrates several simplest tools, each recognizing a particular regulatory site by its textual pattern, into a data mining system rSNP_Guide. For maximizing a user’s trust to rSNP_Guide predictions, we have made all computation results available, thus, all intermediate and final results could be verified by user.

As known, the “false positives” obstacle is now the main limit of the pattern-based site recognition applicability [2]. Table 2 illustrates this drastic limit by the ratio “false positives”/total recognitions, N + , with the values up to 0.35. Nevertheless, one can see that our clustering algorithm, which involves a well known molecular mechanism of the

DNA/protein binding into the in-depth SNP-analysis, has successfully discarded these "false positives".

Since genetic disease may be caused not only by the presence/absence alterations of a site, but also by quantitative alterations of binding efficiency (e.g. erythroid-specific DNA-binding protein affinity alterations cause δ-thalassemia [26]), we are planning to accompany the next version of the rSNP_Guide by our earlier developed Web-tools regressing the activity estimate of a given site upon its sequence [16].

This allows us to conclude that the data mining based on integration of available tools and taking into account the knowledge present in biological sequences gives many perspectives for future in-depth SNP-analysis.

Acknowledgement

The work was supported by grant RFBR-01-04-49860 (Russia).

References

[1] G. Marth, R. Yeh, M. Minton, R. Donaldson, Q. Li, S. Duan, R. Davenport, R. Miller, P.-Y. Kwok, Single-nucleotide polymorphisms in the public domain: how useful are they? Nat. Genet. 27 (2001) 371–372.

[2] E. Uberbacher, Y. Xu, R. Mural, Discovering and understanding genes in human DNA sequence using GRAIL, Methods Enzymol. 266 (1996) 259–281.

[3] G. Vasiliev, V. Merkulov, V. Kobzev, T. Merkulova, M. Ponomarenko, N. Kolchanov, Point mutations within 663-666 bp of intron 6 of the human TDO2 gene, associated with a number of psychiatric disorders, damage the YY-1 transcription factor binding site, FEBS Lett. 462 (1999) 85–88.

[4] M. Andrade, P. Bork, Automated extraction of information in molecular biology, FEBS Lett. 476 (2000) 12–17.

[5] R. Kamimura, S. Bicciato, H. Shimizu, J. Alford, G. Stephanopoulos, Mining of biological data II: assessing data structure and class homogeneity by cluster analysis, Metab. Eng. 2 (2000) 228–238.

[6] R. Kamimura, S. Bicciato, H. Shimizu, J. Alford, G. Stephanopoulos, Mining of biological data I: identifying discriminating features via mean hypothesis testing, Metab. Eng. 2 (2000) 218–227.

[7] D. Scheurle, M. DeYoung, D. Binninger, H. Page, M. Jahanzeb, R. Narayanan, Cancer gene discovery using digital differential display, Cancer Res. 60 (2000) 4037–4043.

[8] P. Bourne, Bioinformatics meets data mining: time to dance? Trends Biotechnol. 18 (2000) 228–230.

[9] P. Hajek, T. Havranek, Mechanizing Hypothesis Formation—Mathematical Foundations for a General Theory, Springer Verlag, Heidelberg, 1978.

[10] P. Fishburn, Utility theory for Decision Making, Jonh Wiley and Sons, New York, 1970.

[11] L. Zadeh, Fuzzi sets, Information and Control 8 (1965) 338–353.

[12] A. Kel, M. Ponomarenko, E. Likhachev, Y. Orlov, I. Ischenko, L. Milanesi, N. Kolchanov, SITEVIDEO: a computer system for functional site analysis and recognition. Investigation of the human splice sites, Comput. Applic. Biosci. 9 (1993) 617–627.

[13] J. Ponomarenko, M. Ponomarenko, A. Frolov, D. Vorobyev, G. Overton, N. Kolchanov, Conformational and physicochemical DNA features specific for transcription factor binding sites, Bioinformatics 15 (1999) 654–668.

[14] M. Ponomarenko, A. Kolchanova, N. Kolchanov, Generating programs for predicting the activity of functional sites, J. Comput. Biol. 4 (1997) 83–90.

[15] M. Ponomarenko, J. Ponomarenko, A. Frolov, O. Podkolodnaya, D. Vorobyev, N. Kolchanov, G. Overton, Oligonucleotide frequency matrices addressed to recognizing functional DNA sites, Bioinformatics 15 (1999) 631–643.

[16] J. Ponomarenko, D. Furman, A. Frolov, N. Podkolodny, G. Orlova, M. Ponomarenko, N. Kolchanov, A. Sarai, ACTIVITY: a database on DNA/RNA sites activity adapted to apply sequence-activity relationships from one system to another, Nucleic Acids Res. 29 (2001) 284–287.

[17] J. Ryan, P. Barker, M. Nesbitt, F. Ruddle, KRAS2 as a genetic marker for lung tumor susceptibility in inbred mice, J. Natl. Cancer. Inst. 79 (1987) 1351–1357.

[18] D. Comings, R. Gade, D. Muhleman, C. Chiu, S. Wu, M. To, M. Spence, G. Dietz, E. Winn-Deen, R. Rosenthal, H. Lesieur, L. Rugle, J. Sverd, L. Ferry, J. Johnson, J. MacMurray, Exon and intron variants in the human tryptophan 2,3-dioxygenase gene: potential association with Tourette syndrome, substance abuse and other disorders, Pharmacogenetics 6 (1996) 307–318.

[19] J. Knight, I. Udalova, A. Hill, B. Greenwood, N. Peshu, K. Marsh, D. Kwiatkowski, A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria, Nat. Genet. 22 (1999) 145–150.

[20] C. Spek, V. Lannoy, F. Lemaigre, G. Rousseau, R. Bertina, P. Reitsma, I. Type, C. protein, deficiency caused by disruption of a hepatocyte nuclear factor (HNF)-6/HNF-1 binding site in the human protein C gene promoter, J. Biol. Chem. 273 (1998) 10168–10173.

[21] L. Ludlow, B. Schick, M. Budarf, D. Driscoll, E. Zackai, A. Cohen, B. Konkle, Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome, J. Biol. Chem. 271 (1996) 22076–22080.

[22] J. Carew, E. Pollak, K. High, K. Bauer, Severe factor VII deficiency due to a mutation disrupting an Sp1 binding site in the factor VII promoter, Blood 92 (1998) 1639–1645.

[23] S. Fucharoen, K. Shimizu, Y. Fukumaki, A novel C–T transition within the distal CCAAT motif of the G gamma-globin gene in the Japanese HPFH: implication of factor binding in elevated fetal globin expression, Nucleic Acids Res. 18 (1990) 5245–5253.

[24] J. Ponomarenko, T. Merkulova, G. Vasiliev, Z. Levashova, G. Orlova, S. Lavryushev, O. Fokin, M. Ponomarenko, A. Frolov, A. Sarai, rSNP_Guide, a database system for analysis of transcription factor binding to target sequences, Nucleic Acids Res. 29 (2001) 312–316.

[25] F. Hampel, E. Ronchetti, P. Rousseeuw, W. Stahel, Robust Statistics—The Approach Based on Influence Functions, Jonh Wiley & Sons, New York, 1986.

[26] P. Moi, G. Loudianos, J. Lavinha, S. Murru, P. Cossu, R. Casu, L. Oggiano, M. Longinotti, A. Cao, M. Pirastu, δ-Thalassemia due to a mutation in an erythroid-specific binding protein sequence $3'$ to the δ-globin gene, Blood 79 (1992) 512–516.