THE METHOD OF IDENTIFICATION OF NUCLEAR RECEPTOR BINDING SITES

[1,3]Cheremushkin E., [1]Cheremushkina E., [2]Kel-Margoulis O., [2]Tchekmenev D., [2]Kel A., [2,4]Wingender E.

[1] **Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia**

[2] **BIOBASE GmbH, Halchtersche Strasse 33, 38304Wolfenbuettel, Germany**

[3] **Institute of Informatics Systems SB RAS, Novosibirsk, Russia**

[4] **University of Gottingen, Germany**

Nuclear receptors (NR) function as ligand-activated transcription factors. They are involved in regulation of reproduction, development, general metabolism and other processes. An inappropriate nuclear receptor signaling is a key pathological determinant in many diseases, such as cancer, obesity, diabetes, cardiovascular disorders. In this work we try to create a method of recognition nuclear receptors sites on DNA.

Data Structure

NR binding site often consists of two hexanucleotide half-sites. Distance between half-sites is variable and sometimes depends on a type of receptor that recognizes the site. Half-sites can be in different orientations to each other. The same half site is very similar for a large number of nuclear receptors. One site can be a target for different nuclear receptors. Therefore, it is very important to identify sites, to understand what nuclear receptor binds to each site.

Recognition Method

Since the site consists of 2 conservative domains, and a distance between them can vary, denote double-core model of recognition M_k: $M_k = \langle m_1, m_2, d_1, d_2 \rangle$, where m_1, m_2 are weight matrices [1] 6 bp. long; d_1, d_2 are minimal and maximal space length respectively between half-sites. Let $w_1(i)$ and $w_2(j)$ are the weights of m_1 and m_2 in positions i and j on the sequence respectively.

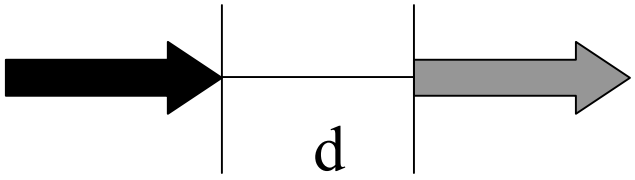

Fig.1*. NR site consists of two domains. and a variable distance between them*

The site will be considered to be recognized, if the weight $w = \frac{w_1(i) + w_2(j)}{2}$ is more than given cutoff and a distance between half-sites $d \in [d_1, d_2]$.

The procedure of NR sites recognition is following: if a typical half-site was recognized on the sequence, we look for potential NR sites by using models $M_1,...,M_K$ where K is a number of all models. In a case of some potential sites for one model M_k we take a site with maximal weight of recognition w_k. If the model M_k is not recognized in this region, assume weight of recognition $w_k = 0$. Then this sites are classified according to the weights w_k by using decision tree method.

Training Method.

The method of finding models M_k: $S = (S_1,...,S_m)$ is a training set of NR site sequences. For each subset $S' = (S_1',...,S_m')$ of set S two collections of subsequences $S^1 = (s_1^1,...,s_n^1)$ and $S^2 = (s_1^2,...,s_n^2)$ are defined, $s_i^1, s_i^2 \in S_i'$, length of $s_i^j = 6$.

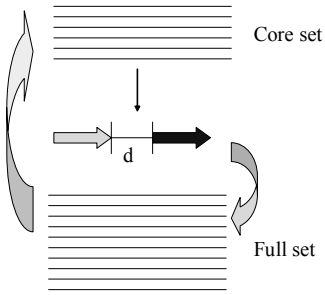

Fig. 2

Applying classical procedure of Gibbs sampling [2] we find S^1 and S^2 such as s_i^1 are similar to each other, и s_i^2 are similar to each other. Matrices m_1 and m_2 can be constructed by using S^1 and S^2 respectively. Than distances $d_1 = \min_i\left(d(s_i^1, s_i^2)\right)$ and $d_2 = \max_i\left(d(s_i^1, s_i^2)\right)$ were calculated. Define an initial subset $S_{[0]}$ named "core set". Than we make a model $M_{[0]}$ and add a sequence from $S \setminus S_{[0]}$ to $S_{[0]}$ for which the weight $w_{[0]}$ of model $M_{[0]}$ is maximal. Consequently, we get a model $M_{[1]}$. We continue a procedure of adding until the weight $w_{[k]}$ is bigger, then initially given cutoff C. (Fig. 2).

After the end of the procedure we get a model *M* that describes a set *S*. Hence we get different models $M_1,...,M_T$ for different classes of NR sites.

After that we use for all learning sets of sites S_{all} a procedure of searching described before. For every sequence *s*, where a typical site is recognized, we get a vector $V_s = (V_1,...,V_T)$ of model weights $M_1,...,M_T$. And according to this set of vectors we build a decision tree by the method *KART*-style. As a criterion of choice we take a minimal mistake of cross-validation. The cross-validation is maid after each step of building the tree. A part of learning set is cutting out and serve as control set. Then the procedure is repeated N times and a mistake of cross-validation averages over all tests.

Data

There are a number of NR binding sites, used for learning sets in the table. To create core sets there were taken 5-6 sites from each group, that are often used in research of this NR.

The sequences of sites were taken from TRANSFAC database (3). Some sites can bind to more than one type of NR, therefore a number of sites in groups is not the same as a amount of sites for all NRs.

Table 1. A number of NR binding sites, used for the analysis

NR	Number of sites	Group	Number of sites
AR	30	Group steroid	138
ER	45		
GR	59		
PR	27		
T3R	17	Group 1 thyroid	48
RAR	28		
LXR	15	Group 2 thyroid	65
VDR	26		
PXR	15		
FXR+RXR	10		
SF1	25		
COUP	21	Group 3 thyroid	61
PPAR+RXR	25		
HNF4	42		

Results.

There was built double-core model for each factor from table 1, for 4 groups of factors (table 1) and for 10 different repeats. Resulting decision tree has 18 nodes and 19 leaves. Also it was made a program for recognition of NR sites.

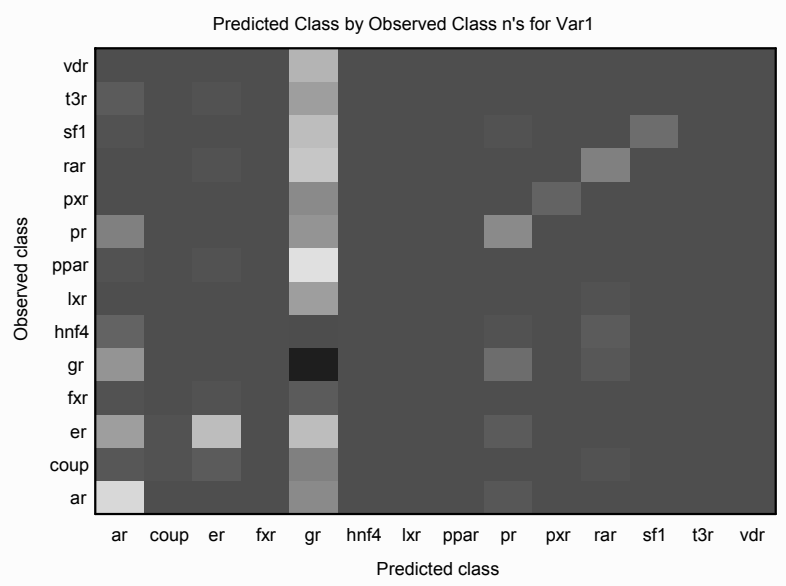

Fig .3.

atattcctgg tcagcgtg**ac cggagctgaa ag**aacacatt
gatcccgtga

0.043	0	0	0.2	0	0.1	0	0	0	0	0	0	0.2	0
AR	CAR	COUP	ER	FXR	GR	HNF	PPA	PR	PXR	RAR	SF1	T3R	VDR

Fig. 4

Literature

1. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E.
2. Lawrence, C.E., Altschul, S.F., Bogouski, M.S., Liu, J.S., Neuwald, A.F., and Wooten, J.C. (1993), "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment," Science, 262, 208-214.
3. TRANSFAC database as a bridge between sequence data libraries and biological function. Wingender E, Karas H, Knuppel R.
4. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.