PREDICTION OF INTERFERON-INDUCIBLE GENES IN HUMAN GENOME

Ananko E.A.*, Kondrakhin Yu.V., Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
*Corresponding author: e-mail: eananko@bionet.nsc.ru

Key words: recognition of transcription factor binding sites, interferon-stimulated genes, genome annotation

SUMMARY

Motivation: Application of the methods of computer-assisted genome annotation coupled with large-scale experimental studies may be helpful in determining possible functions of numerous unstudied genes. The search for interferon-inducible genes is of particular interest. As known, interferons modulate the work of the immune system: they exert antiviral, antibacterial, and antitumoral effect. Although the system of interferons is being actively studied during several dozens years, the mechanisms of its functioning are still not clear in many respects.

Results: By using the methods developed for recognition of interferon-inducible genes, an analysis of DNA sequences of more than 2000 genes within the length limits from -1000 to +1000 bp relatively transcription start was performed. We have detected 78 genes that could be interferon-inducible with high probability and could participate in supporting some interferon’s functions.

Availability: The list of predicted interferon-inducible human genes obtained in the course of the work considered is available at http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/iig-trrd/ISG_predicted.html.

INTRODUCTION

Interferons (IFNs) are classified into two types: IFNs type I, or virus-inducible acid-resistant interferons (e.g., leukocyte IFN-α, fibroblast IFN-β, IFN-δ, IFN-ε, IFN-κ, IFN-ω, IFN-τ, and IFN-ζ) and IFN type II, or immune acid-liable IFN-γ. Type I interferons are mainly support antiviral state of the organism (Pestka et al., 2004), whereas IFN-γ makes larger impact in providing antibacterial and antiparasitic responses (Decker et al., 2002). Also, IFN-γ was shown to participate in development of autoimmune states (Baccala et al., 2005).

By studying signal transduction pathways of interferon system, it was established that IFNs type I cause activation of ISGF3 transcription factor, whereas IFN-γ – activation of the STAT1 homodimer (Platanias, 2005). Type I interferons activate also Akt serine-threonine kinase and p38/MAP-kinase cascades, as well as the signal transduction pathways leading to activation of NF-κB and p53 transcription factors: all these factors participate in the antiviral immune response and tumor suppression (Pestka et al., 2004;
Interaction of these and some other transcription factors with transcription factor binding sites in regulatory regions of interferon-stimulated genes (ISG) mediate significant increase of gene transcription.

**METHODS AND ALGORITHMS**

For recognition of transcription factor binding sites, we have used an additive recognition function with application of statistic simulation approach (Kondrakhin et al., 2006).

For selecting individual sites for recognition of ISG, we have applied a statistical test that compares two binomial values, that is, mainly those sites were selected, relative occurrence frequency of which within the interferon-inducible genes (F1) was statistically increased in comparison to that of the genes entering the control sample of genes extracted from the EPD database (F2). To select a pair of sites, we have used the standard statistical test $\chi^2$. For majority of selected sites and pairs of sites, p-value, was $< 0.001$.

The score at each position of arbitrarily chosen sequence SEQ0 was calculated as follows. First, we have selected $m$ objects from the training sample (i.e., sites and pairs of sites), $T_1$, $T_2$, ..., $T_m$. Then for the $i$-th position of the sequence SEQ0, we calculate $m$ of weights $w_1, w_2, ..., w_m$ by the formula:

$$w_i = \begin{cases} 1, & \text{if the i-th object, } T_i, \text{ is not found at respective positions of the sequence SEQ0;} \\ \frac{F1}{F2}, & \text{in case the i-th object, } T_i, \text{ is found at respective positions of the sequence SEQ0}. \end{cases}$$

Then we calculate the score by multiplying the weights $w_1, w_2, ..., w_m$, so that

$$\text{SCORE} = w_1 \cdot w_2 \cdot ... \cdot w_m. \quad (1)$$

Then we calculate the score for all positions of the sequence SEQ0 and select the position with the maximal score.

Notably, the more is the number of the objects selected at relevant positions of the sequence SEQ0, the higher is the score. In other words, SCORE is a function measuring similarity between the sequence SEQ0 studied and the training sample, out of which the objects $T_1$, $T_2$, ..., $T_m$ were extracted.

For calculation of SCORE for each method (see results), the same multiplicative function (1) was applied, but for each method its own set of pre-selected objects (i.e., sites and pairs of sites), $T_1$, $T_2$, ..., $T_m$, was used.

**RESULTS**

By using three methods of ISG recognition (Kondrakhin et al., 2006), we have studied 1664 human genes, within the regions from -1000 to +1000 bp relatively transcription start site, annotated in the EPD database. In order to minimize type II error, we have ordered very stringent threshold limits for all these three methods applied simultaneously. The threshold value of recognition function for the method 0 (induction by any IFN) equals to 0.4. The values of the other two recognition functions should also exceed the threshold level equaling to 0.4 for the method 1 (induction by type I IFNs) and 0.3 for the method 2 (induction by type II IFN).

The verification of recognition methods was accomplished by using the sample of ISG that were identified by microarray data (sample M0 for the positive control). In Table 1, the results of recognition of IFN-inducible genes in different samples of genes are given. In addition to the training sample ISG-TRRD that was compiled on the basis of the TRRD database (Kolchanov et al., 2002), and the sample M0 for the positive control, we have also tested the sample compiled on the basis of EPD database (1664 human genes). As the negative control, we have analyzed two samples containing very small percentage of ISG.
i.e., genes regulated by glucocorticoids (GR-TRRD) and genes of lipid metabolism (LM-TRRD). Recognition was performed under the same conditions for all the samples: the sequences from -1000 to +1000 relatively transcription start site were analyzed.

In total, among 1664 human genes extracted from the EPD database, we have found 78 genes that potentially response to stimulation by interferons (Table 1). Four genes out of 78 were previously included into the training sample. For 60 genes detected, the stimulation by interferons was not reported yet. In addition, for 13 genes, experimental evidence was obtained demonstrating that transcription is enhanced under the action of interferons by means of RNA microarray data. In 28 genes found, the regions of maximal sensitivity to interferon induction were located in promoter region (from -200 to +50 bp relatively transcription start).

Table 1. Recognition of interferon-inducible genes among different samples under the threshold limitations equaling to 0.4 for the method 0 and method 1, and 0.3 for the method 2

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample size (total number of sequences)</th>
<th>Genes recognized (total number)</th>
<th>Genes recognized (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISG-TRRD</td>
<td>72</td>
<td>17</td>
<td>23.6</td>
</tr>
<tr>
<td>M0</td>
<td>1005</td>
<td>156</td>
<td>15.5</td>
</tr>
<tr>
<td>EPD</td>
<td>1664</td>
<td>78</td>
<td>4.7</td>
</tr>
<tr>
<td>GR-TRRD</td>
<td>70</td>
<td>1</td>
<td>1.7</td>
</tr>
<tr>
<td>LM-TRRD</td>
<td>58</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

DISCUSSION

The potential ISGs found may be classified into several functional groups with respect to biological activities of interferons as the genes involved into (i) immune and inflammatory response, (ii) regulation of cell proliferation and differentiation, and (iii) antitumoral effect (see the complete list of genes at http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/iig-trrd/ISG_predicted.html). For 21 genes, it was difficult to relate their activity with biological function of interferons, so they are considered as possible over-estimation.

Due to our estimates, by taking into account possible over-estimation (in total, 21 genes out of 78 recognized, or 1 % of the sample), human genome carries about 3000 ISG. This value does not contradict to microarray data. For example, only in the primary culture of monocytes isolated from peripheral blood of patients diseased by hepatitis C, at least two-fold induction of 1012 genes was registered under the action of IFN-α during 6 hours after simulation (Ji et al., 2003), whereas in IFN-γ-stimulated macrophages, 632 genes were induced (Ehrt et al., 2001). In hepatocarcinoma cell line HepG2, out of 14 112 genes considered, more than 400 genes were induced by two-fold by IFN-α and 405 genes were induced by IFN-γ (Xiong et al., 2003).

Simultaneous application of computer-assisted methods for recognition of genes simulated by various IFNs enables to reveal in mammalian genome with high accuracy ISG that are involved in interferon system functioning.

ACKNOWLEDGEMENTS

The authors are grateful to E.V. Ignatieva for kindly providing the sample of lipid metabolism genes and to G.V. Orlova for translating the manuscript from Russian into English. The work was supported by the Russian Government (Contracts Nos 02.434.11.3004, 02.467.11.1005) and Siberian Branch of the Russian Academy of
REFERENCES


