Structure of the promoter region in NF-kappaB dependent genes in view of NF-kappaB transcription factors family

Marta Iwanaszko, *Silesian University of Technology*

Abstract

The NF-ϰB family plays a prominent role in the innate immune response, cell cycle activation or cell apoptosis. Upon stimulation by pathogen-associated patterns, such as viral RNA a kinase cascade is activated, which strips the NF-ϰB of its inhibitor IϰBα molecule and allows it to translocate into the nucleus. Once in the nucleus, it activates transcription of approximately 90 genes whose kinetics of expression differ relative to when NF-ϰB translocates into the nucleus, referred to as Early, Middle and Late genes. It is not obvious what mechanism is responsible for segregation of the genes' timing of transcriptional response.

It is likely that the differences in timing are due, in part, to the number and type of transcription factor binding sites (TFBS), required for NF-ϰB itself as well as for the putative cofactors, in the Early vs Late genes. We therefore applied an evolutionary analysis of conserved TFBS. We found that Early genes had significance enrichment of NF-ϰB binding sites occurring in evolutionarily conserved domains vs. genes in the Late group. The similarities observed among the Early genes were seen in comparison with distant species, while the Late genes promoter regions were much more diversified.

1. Background

NF-ϰB is a family of transcription factors (1) that plays a prominent role in innate immune response among other cellular processes, as reviewed in Tian et al. (2005). Upon stimulation by pathogen-associated molecular patterns, such as viral RNA, a kinase cascade is activated, which eventually strips the NF-ϰB of its inhibitor IϰBα molecule and allows it to translocate into the nucleus. In the nucleus, NF-ϰB binds to specific palindromic sequences in the regulatory sequences of promoters to activate the transcription of a number of genes (approximately 90, of which 74 were systematically examined, Tian et al. 2005). The dynamics of NF-ϰB translocation has been studied both experimentally and using mathematical and computer modelling (3-5). Inspection of the mRNA transcript profiles has further shown that the NF-ϰB-dependent genes can be categorized by the timing of their activation relative to NF-ϰB's translocation into the nucleus (2). Notably, the Early genes' peak response occurs at about 30-60 min. after NF-ϰB translocation, as opposed to the Middle genes' response at about 3 hrs. and the Late genes' response at up to 6 hrs. Interestingly, these categories encode distinct molecular functions, the Early genes being predominantly cytokines, Late genes encoding cell surface adhesion molecules and signalling adapter molecules and Middle genes overlapping Late genes' functions in control of signalling molecules and expression of cell-surface receptors.

It is not obvious what mechanism is responsible for segregation of the genes' timing of transcriptional response. One likely hypothesis might be that the later the gene is the more cofactors are required to activate it. This hypothesis gave rise to a mathematical model in Paszek et al. (6). Another hypothesis is that NF-ϰB has to be primed by a post – translational modification such as amino acid – specific phosphorylation or acetylation to act as a transcription factor for a given gene, and that such processing requires additional time in some cases. This latter hypothesis was at least in part confirmed by Nowak et al. (7).

The question we address here is how gene's expression is regulated by transcription factor binding sites (TFBS) in the gene's promoter. NF-ϰB family is sequence specific, with four identified binding motifs corresponding to different family members. Identified binding motifs are 10 nucleotides long (except for the motive for the heterodimeric particle) and have a characteristic guanine triplet (GGG) opening the motif. Given this, it seems correct to use the software finding TFBS in genetic sequences. Advances in gene expression analysis technologies allow for detection using computational TFBS methods and development of databases containing position weight matrices (PWMs), such as JASPAR (8) and TRANSFAC (9). Analysis of sequences for the presence of known TFBS only by using PWMs can produce a large number of false positive predictions; therefore computational TFBS detection must be enriched with some other methods helping find functionally relevant TFBS (10). This can be accomplished using phylogenetic footprinting, which is based on the assumption that TFBS should be highly conserved in comparison to non – regulatory regions close to genes (11). This approach is used by ConSite (12), which uses ORCA algorithm (13) for phylogenetic footprinting and JASPAR database for TFBS sequence identification. Recent research suggests that in transcriptional regulatory regions modules occur, which contain clusters of TFBS (14) that can be distinguished from non – regulatory areas by high conservation.

Tab. 1

Transcription factors association with gene data set

Rank	Matrix	Transcription Factor	Association Score	P-Value
1	NFKAPPAB65_01	Rela	15.083	0.00e+00
2	NFKB_Q6_01	Nf-kappab1 , Nf-kappab2	13.488	0.00e+00
3	NFKAPPAB_01	Rela	13.336	0.00e+00
4	CREL_01	C-rel	12.883	0.00e+00
5	NFKB_Q6	N/A	11.200	0.00e+00
6	NFKB_C	N/A	9.671	0.00e+00
7	NFKAPPAB50_01	N/A	7.199	0.00e+00
8	CDPCR1_01	Cutl1	5.367	8.30e-05
9	CDPCR3HD_01	Cutl1	5.265	1.03e-04
10	PAX2_01	Pax-2 , Pax-2a	4.845	2.66e-04

2. Results

To determine if there exists regulatory association between NF-ϰB family TFs and our dataset we used PASTAA software (15). Results for top 10 associated TFs are presented in Table 1. As we assumed the highest affinity for gene set exists for NF-ϰB family members. The basic descriptive statistics of the number of binding motifs found in the study, are collected in Table 2, which is listing group – by – group (rows) the number of motifs found (columns), for NF-ϰB – related genes itemized and for other sequences (random sequences and shuffled real promoter sequences) jointly. This study revealed that among chosen NF-ϰB-dependent genes, the average number of separated NF-ϰB-family TFBS detected in dataset equal to 6.07 per sequence, while the number in random sequences and shuffled sequences is about 2

TFBS. This comparison indicates that there exists a substantial difference between occurrence of NF-ϰB-related TFBS not only between NF-ϰB – dependent and random sequences, but also among the Early and Middle versus the Late groups. There is a considerably high percentage of NF-ϰB-related TFBS (multiple and overlapping) among the Early and Middle genes, in contrast to a lower number of TFBS found in the promoters of Late genes. Wilcoxon test of abundance shows that there is a statistically significant difference (at the significance level $\alpha = 0.05$) between the randomly generated and shuffled real promoter sequences and the promoter sequences of NF-ϰB-dependent genes (Early: $p=1{,}58^{-5}$, Middle $p=2.8^{-6}$, Late $p=1.6^{-4}$).

2.1 Eevolutionary conservation of TFBS in promoter regions

Cross – species comparison revealed that conservation of NF-ϰB family - related TFBS motifs is much higher in the Early genes group than in the Late genes group. The highest numbers of common DNA binding motifs considered were found in the locations where the adjusted promoter sequences were highly conserved. For all Early genes, NF-ϰB-family related TFBS motifs conserved between most pairs of species (the exception being TNF between mouse and cow) were found. As we presumed the best promoter sequence conservation and interspecies conservation of TF binding motifs persists between human and chimp, followed in many cases by that between human and cow. In the case of two Early genes, *REL* and *TNFAIP3* comparison, no conserved NF-ϰB-family related TFBS were found between chimp or mouse and cow. In human versus cow comparison two single non-overlapping binding sites were found, but this is a low score in comparison with the number of conserved TFBS found in other Early genes. In human versus chimp comparison in nearly all Early genes, all NF-ϰB-family related TFBS found were conserved. Only in the case of the *IϰBa* gene the number of conserved TFBS is lower than the number of TFBS found separately for each of these species. The likely cause of this difference in promoter sequence is a long shift in promoter sequence alignment. Comparing given promoter sequences we can observe that in the case of the *IϰBa* gene in human, the groups of TFBS found are located in the distant region of the promoter whereas in other studied species, NF-ϰB family – related TFBS are located in the proximal region of the promoter, and mostly conserved.

Tab. 2

Participation of NF-κB – family binding motifs among human TFBS vs. random sequences

Group of genes	Number of TFBS found				Avg. number of TFBS
	NFkappaB	c-Rel	p50	p65	
Early	31	27	18	17	
Middle	32	34	16	20	
Late	18	22	14	12	
Sum for dataset	81	83	48	49	**6,07**
Sum for 50 random sequences	28	49	15	26	**2,36**
Average sum for shuffled sequences	22,6	36,7	13	15,5	**2,09**

In the case of the Middle genes group, the highest number of conserved NF-ϰB-family related TFBS is found in the *RELB* gene. Among all species comparisons, conservation of this gene's promoter sequences reaches 90% and more in the proximal region. Because of this, the most abundant multiple TFBS located close to the coding region are well conserved among all species. In the other two Middle genes, conservation of sequences is weaker and accordingly the conserved NF-ϰB family – related TFBS are less numerous. Even if structure of TFBS looks similar we discover lower accuracy in sequences. In the *NFKB1* gene a similar arrangement of binding sites along the promoter sequence can be observed, and many NF-ϰB-family related TFBS were found in single promoter analysis, but cross – species comparison reveals quite low conservation of promoter sequences and visible differences between human, chimp and mouse versus cow promoter structures. No NF-ϰB family – related TFBS were found in human versus mouse and human versus cow comparison due to very low overall promoter sequence conservation. However, chimp versus cow comparison revealed common multiple NF-ϰB family – related TFBS. The worst overall conservation in the Middle genes is observed in *TRAF2*, where for only one comparison, human versus chimp, one unique NF-ϰB-family related TFBS was found.

In the Late genes group cross – species comparisons, the lowest numbers of common conserved NF-ϰB family – related TFBS were found. This study revealed a great divergence between the promoter structures among considered species. Moreover there are only two genes, NFϰB2 and *TNIP1* in which the conserved NF-ϰB-family related TFBS can be found among all species. The best conservation results are between human and chimp and between human and cow. In other Late genes, usually only one or two cross-species comparisons reveal any existing common TFBS. Study of the Late genes group show that, if an orthologue gene exists, cow promoter regions are filled with overall greater number of NF-ϰB family - related TFBS than in other species (*ICAM1, NFϰB2, TRAF1, PTGES*). In human and chimp *TRAF1* gene there is no NF-ϰB family – related TFBS and no conserved motifs were found between other species. In *TRAF3* there is a great similarity between human and mouse promoter, while a very low one between human and chimp or human and cow. Overall, the NF-ϰB family – related TFBS along the Late genes promoters sequences are distributed more sparsely than in the Early genes. TFBS are more scattered in promoter region and less multiple/overlapping TFBS are found in comparison with the Early genes. The degree of sequence conservation between the Late genes in pairs of species also differs from that in Early genes, in most cases not exceeding 50%.

3. Conclusions

Comparison of the Early, Middle and Late genes groups reveals that strongest similarities among species can be found in the Early genes promoters. The Early genes group has the highest conservation percentage of promoter sequences and good sequence alignments, which is the cause of very good cross – species conservation of NF-ϰB-family related TF binding motifs. During evolution these non – coding sequences have maintained a very similar structure, which can serve as a proof of important regulatory functions concerning gene expression that have not changed significantly even tens of millions of years since the divergence from a common ancestor. Moreover, this may suggest that the regulation pattern of these group of genes may have an effect on the result of gene expression and may be more universal, therefore likely to be shared between other species not included in this study.

Analyses of the Late genes expose significant differences in the promoter structure, number and location of NF-ϰB-family related TFBS in promoter sequence and a low number, if any, of conserved NF-ϰB-family related TFBS. This suggests that during evolution, promoter sequences of Late genes became more species – specific and the way that regulation of gene expression is accomplished, has been relatively quickly changing, with increasing evolutionary distance. Comparing human and chimp promoters sequences with those of cow and mouse, suggests that in the case of Late genes some NF-ϰB-family related TFBS lost their functionality and were abandoned or reorganized during species evolution.

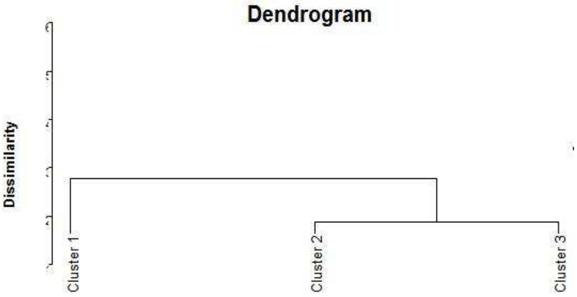

Fig 1. Hierarchical clustering of Middle genes

Tab.3
Hierarchical clustering of Middle genes

Cluster	1	2	3
Size	3	10	5
	KLRC3	TNFAIP2	RELB
	SLC7A2	SOD2	BCL3
	GCH1	BIRC2	CFB
		BID	GFPT2
		NFKBIE	NFKB1
		IFNGR2	
		ECE1	
		CD83	
		TRAF2	
		SDC4	

We noted that the Middle genes can be distinguished into two groups:, 1. Early-like where the promoter region contain relatively high number and clustering of NF-ϰB-family TFBS like the Early genes, and 2. the Late-like genes, which have low number and no significant clustering of NF-ϰB – family TFBS in their promoters like the Late genes. Inspection of the hierarchical clustering patterns even shows 3 groups, but two are more related to each other than to the last one (Figure 2, Table 3), which shows that Middle gene promoter regions are relatively rich in NF-ϰB-family binding motifs compared to Early genes.

In non-NF-ϰB – dependent genes, NF-ϰB – family related TFBS generally represented less than 2% of all TFBS and were single, non-clustered sites. It has been suggested by Wunderlich and Mirny (14) based on information-theory considerations that sites of such structure are non-functional, since non-clustered binding sites may not be recognizable by the corresponding TFs. In the NF-ϰB – dependent genes and particularly in the Early and the Middle groups, the NF-ϰB family – related TFBS are much more numerous (around 10%) and are usually clustered together.

3.1 Discussion

Our study revealed that promoter structures in the Early, Middle and Late genes are different but conserved in four species. The Early genes were found to have more NF-ϰB-family related TFBS, among which most are multiple and located close to each other, compared to genes belonging to the Late group. The cofactor hypothesis (6) may not be a sufficient explanation of such difference in expression timing between the Early, Middle and Late genes. While eukaryotic TFs have low specificity, and are not as precisely targeted to functional cognate sites as prokaryotic TFs are (14), there is a possibility that, shortly after NF-ϰB-family related TFs are released in the nucleus, they are unable to locate functional binding sites in genes classified as Late. Because of that low specificity NF-ϰB family – related TFs may bind to non-functional sites, delaying the time of gene expression. Without signal about expression of gene, further TF molecules continue to bind to the available and recognizable binding motifs. Somewhat similarly as in Paszek et al. (6) cofactor hypothesis, the Late genes may require more than one functional binding site to be bound by TF to start expression, and the higher chance of TFs binding to non – functional sites is the reason that a longer time is required before TF reaches the functional cognate sites.

Extending the gene dataset to more NF-ϰB – dependent and interferon dependent genes, may provide more knowledge about gene expression mechanisms and control innate immunity.

5. Methods and data

5.1. Selection of species and genes

For this research four mammalian species were chosen with their evolutionary distance from humans as the main guideline. Chimpanzee, was chosen as the closest to human, to inspect hypothetically most recent changes in TFBS conservation. Mouse was chosen because of many similarities, good genome representation in bases and proper evolutionary distance, which may show well conserved traits in cross species comparison with human genome. Cow was chosen as the most distant from human in this comparison; however it still maintains many similarities to the human sequences, some of which are better than in mouse.

Selection of genes in dataset is from Tian et al. work (2) and their division based on dynamics of transcriptional response.

5.2. Retrieval of sequences and databases employed

UCSC Genome Browser was used to retrieve human 1000 bp 5' sequences to the first exon, which we call here promoter sequences. Promoter regions for other analysed species were retrieved by BLATing human sequence to other species genome and by analysing synteny blocks. If no significant match was found, then sequence 1000 bp upstream from TSS was assumed to represent promoter region for certain gene. Ensembl database was employed for acquisition of some of chimpanzee and cow genes. Profiles of chosen TFs were drawn from the JASPAR database and then converted to log – scaled position weight matrices (PWMs) in order to evaluate possible binding sites in the input sequence (12).

5.3. Bioinformatic tools

Analyses of promoter sequences and cross - species comparisons were conducted mostly using ConSite, rVista and NucleoSeq (12, 13, 16). Analysis concerning abundancy of NF-ϰB family motifs was done using PASTAA (15) and scripts for Matlab and MS Excel. Hierarchical clustering was implemented in Excel and Matlab.

4. Bibliography and authors

4.1. Bibliography

[1] Baeuerle PA, Baltimore D.: *NF-ϰB: Ten years after*. Cell 1996 (87): 13

[2] Tian B, Nowak DE, Brasier AR *A TNF-induced gene expression program under oscillatory NF-ϰB control;* BMC Genomics 2005 (6): 137

[3] Hoffmann A, Levchenko A, Scott ML, Baltimore D: *The IkappaB–NF-kappaB signaling module: temporal control and selective gene activation.* Science 2002 (298): 1241

[4] Lipniacki T, Paszek P, Brasier AR, Luxon B, Kimmel M: *Mathematical model of NF-kappaB regulatory module.* J Theor Biol 2004, 228: 195

[5] Cheong R, Hoffmann A, Levchenko A: *Understanding NF-ϰB signaling via mathematical modeling.* Molecular Systems Bio. 2008, 4: 192

[6] Paszek P, Lipniacki T, Brasier AR, Tian B, Nowak DE, Kimmel M*: Stochastic effects of multiple regulators on expression profiles in eukaryotes.* J. of Theoretical Biol. 2005, 233: 423

[7] Nowak DE, Tian B, Jamaluddin M, Boldogh I, Vergara LA, Choudhary S, Brasier AR: *RelA Ser276 Phosphorylation Is Required for Activation of a Subset of NF-ϰB-Dependent Genes by Recruiting Cyclin-Dependent Kinase 9/Cyclin T1 Complexes;* Mol Cell Biol. 2008 28(11): 3623

[8] Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, LenhardB, Sandelin A: *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucl. Acids Res. 2008; 36(Database issue):D102-6

[9] Wingender E, Dietze P, Karas H, Knűppel R: *TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites.* Nucl. Acids Res. 1996, 24 (1): 238

[10] Qiu P: *Recent advances in computational promoter analysis in understanding the transcriptional regulatory network Biochem.* Biophys. Res. Commun. 2003, 309: 495

[11] Hardison R: *Conserved noncoding sequences are reliable guides to regulatory elements.* Trends in Genetics 2000, 16: 369

[12] Sandelin A, Wasserman WW, Lenhard B: *ConSite: web-based prediction of regulatory elements using cross-species comparison.* Nucl. Acids Res. 2004, 32: w249

[13] Portales – Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW: *The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences.* Nucl. Acids Res.2008, 37 (Database issue): D54–D60

[14] Wunderlich Z, Mirny LA: *Different gene regulation strategies revealed by analysis of binding motifs.* Trends Genet. 2009, 25(10); 434

[15] Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA: *PASTAA: identifying transcription factors associated with sets of co-regulated genes.* Bioinformatics 2009, 25 (4): 435

[16] Jaksik R, Rzeszowska-Wolny J**:** *Using NucleoSeq application to identify patterns in m*RNA *nucleotide sequences.* Acta Biochim Pol. 2010

4.2. Authors:

MSc. Marta Iwanaszko

Silesian University of Technology
ul. Akademicka 16
44-100 Gliwice
tel. (032) 237 21 19

email: *marta.iwanaszko@polsl.pl*