COMBINATORIAL TOOLS FOR THE ANALYSIS OF TRANSCRIPTIONAL REGULATION

A. BERGERON, E. GAUL
LACIM, Université du Québec à Montréal
C.P.8888 Centre-Ville, Montréal, Canada, H3C 3P8
{*anne, gaul*}*@lacim.uqam.ca*

D. BERGERON
Département de microbiologie et immunologie, Université de Montréal
C.P. 6128 Centre-Ville, Montréal, Canada, H3C 3J7
bergerdo@ere.umontreal.ca

In this paper, we discuss virtual experiments for the study of major regulatory processes such as translation, signalization or transcription pathways. An essential part of these processes is the formation of protein clusters held together by a small number of binding domains that can be shared by many different proteins. Analysis of these clusters is complicated by the vast number of different arrangements of proteins that can trigger a specific reaction. We propose combinatorial tools that can help predict the effects on the rate of transcription of either changes in transcriptional factors concentration, or due to the introduction of chimeras combining domains not usually present on a protein.

1 Introduction

Most regulatory proteins consist of a combination of discrete modules or *domains* which have been shuffled during evolution[1,2]. Many of these domains are involved in protein-protein interactions modulating various processes within the cell, including gene expression. A domain is a specific region of the protein surface, created by the folds of an amino acid chain. An interaction consists essentially in the transient apposition of two complementary domains. Structural data show that the number of different domains is relatively small, but the great diversity of protein function is partly due to the vast number of arrangements generated by these basic domains.

Along the multiple folds of their polypeptide, a protein harbors several distinct interaction domains. Theoretically, a protein could thus be involved in the formation of clusters composed of several interaction partners. A large number of regulatory proteins, such as transcriptional factors, do not have intrinsic enzymatic activity, but modulate cellular processes by their involvement in the formation of intermolecular clusters central to specific enzymatic activity. The function of these proteins could thus be explained by their ability to drive cluster formation.

In this paper, we present a combinatorial and probabilistic study of the various clusters that can be formed through protein-protein or protein-DNA interactions. We discuss computational experiments based on this model that attempt to replicate experimental findings on human Ewing's tumors. In this case, the accidental exchange of two DNA strands results in the expression of a chimeric protein that combines domains not normally present in a transcriptional factor. In vitro, this protein has been shown to activate transcription at a rate 30 times fold the rate observed in natural conditions, leading to tumoral growth [3].

The proposed computational tools can be helpful in rapidly conducting virtual experiments with new or suspected regulatory proteins. The program can readily adapt to the introduction of new proteins, even when their structure is only partially known. However, the predicting value of the model will depend on networking with databases and other analytical software [4,5] which will provide growing information on the nature of protein domains, their location, and their binding properties.

2 Biological Background

2.1 Signaling Pathways

In response to external stimulus, like growth hormones or cell contacts, biochemical cascades are initiated in the cell, ultimately resulting in changes in gene expression. These signal pathways involve several steps in which specific proteins - the *targets* - are chemically altered, due to the action of kinases or phosphates, for example. The study of molecular mechanisms involved in signaling pathways has shown that each step of the pathway is characterized by the formation of transient protein clusters, necessary to bring the appropriate presented substrat near the proper enzymatic activity, by the intermediate of molecular glue [6]. The phosphorylation state of the target protein in turn determines its capacity of forming subsequent clusters [7] (see Figure 1).

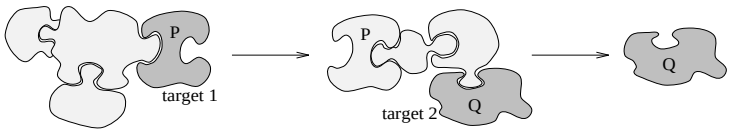


Figure 1: Signaling Pathway

The final step of a signaling pathway is generally connected with the phosphorylation of a transcriptional factor [8], thus modulating gene expression.

2.2 Transcriptional Regulation

A gene, coded by a strand of DNA, contains the necessary information for the synthesis of a protein. The actual synthesis, *gene expression,* is initiated by a process called *transcription* in which gene information is transferred to mRNA. One particular enzyme (DNA-dependent RNA polymerase II, denoted as Pol II) is responsible of the transfer, and must be correctly positioned along the gene in order to initiate transcription. Transcriptional regulation refers to the various mechanisms used by the cell to control the rate at which transcription occurs.

The recruitment and positioning of Pol II is done by clusters of proteins - called *transcriptional factors* - that are formed on a promoter-enhancer section of DNA preceding the actual gene (Figure 2). Each protein in a cluster has a set of *binding domains* on its surface. A particular domain will bind the corresponding domain of another protein or of DNA, allowing the cluster to grow in a Lego-like fashion, and occasionally recruiting Pol II. If Pol II is readily recruited, gene expression is activated [9].

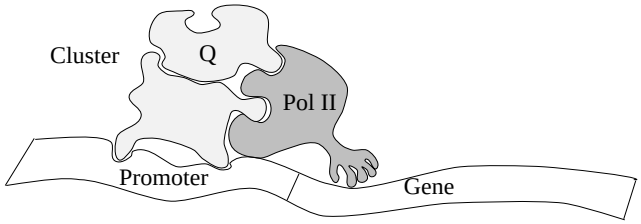


Figure 2: Promoter, Cluster and Gene

In this model, and as observed in molecular processes of signaling pathways, the majority of transcriptional regulators act simply as linkers in the cluster. A given binding domain is not specific to a transcriptional factor. In fact, the number of different domain families is relatively small - see Table 1-, and domains with similar binding properties can be present on several different proteins [1,2,9]. Binding domains can thus be viewed as connectors that can build many different clusters. This combinatorial aspect of cluster formation complicates the prediction of protein function in transcription.

The composition and size of a cluster able to recruit Pol II are predicted to affect the level of gene expression: a higher complexity is expected to correlate with a reduction of expression. Activation will also depend on both the structure of the promoter, and the nature and concentration of the proteins available in the environment - called here a *medium.*

2.3 Molecular Cluster Formation

Clusters are formed by the aggregation of proteins through *interaction* or *binding domains*[9,10]. These are amino acid residue portions - often, but not necessarily a consecutive sequence - that configures into a unique three dimensional shape displaying specific charged and/or hydrophobic groups[11]. Two domains with the same binding properties, but not necessarily the same composition, are considered of the same *type.*

A basic condition for formation of a cluster is that each type of domain t (ex. a basic domain) has one - or more - complementary type t' (ex. an acidic domain) such that there is a non-covalent attraction between t and t' (ex. basic:acidic) encouraging the formation of clusters. A similar kind of binding occurs between protein domains that recognize small specific sequences of DNA,like the TATA-binding factor[9]. By analogy, we will also call binding domains the specific regions of a promoter that are involved in these interactions.

3 The Combinatorial Model

The formal definition of molecular clusters will suppose that some properties of actual clusters are abstracted. In this first attempt, we model the interconnections of the different proteins involved in a cluster. Our goal is to give a precise meaning to *"the ability of a medium to form clusters that will recruit a given protein P".* This notion will be cast as the mathematical expectation of a random variable defined on a population of clusters.

3.1 Basic Definitions

Let T be a finite set of *types of domains.* The *binding relation* $\mathcal{B}$ is a symmetric relation on $T \times T$. This relation stores the information on which pairs of types are complementary. A *protein* is a pair $P = (Ident, Domains)$, where $Ident$ is the name of the protein, and $Domains$ is the multi-set of its domains, consisting of elements of T with their multiplicities. Figure 3 shows an example of a protein with three domains.

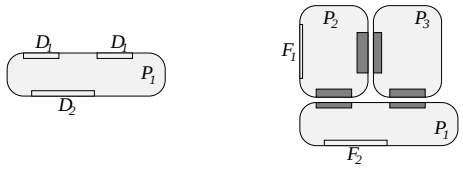


Figure 3: The Protein $P_1 = (Ident, \{2 \times D_1, D_2\})$ and the Cluster $(\{F_1, F_2\}, \{P_1, P_2, P_3\})$

A *cluster* is given by a pair $C = (Free, Members)$, where $Free$ is the multi-set of its free domains, and $Members$ is the multi-set of proteins that are members of the cluster (Figure 3). Note that a protein can be present in several copies in a cluster. A cluster with free domains, but no member, is a *promoter*.

3.2 Recruitment

A *link* is a pair of domains (D, E) such that the types of D and E are complementary, ie. the pair (D, E) belongs to $\mathcal{B}$. The *first projection* of the link (D, E) is D, and its *second projection* is E. Given a cluster C and a protein P, we will say that C *recruits* P *with links* $L_1, \ldots, L_k$ *yielding* C', noted

$$C \xrightarrow[L_1,\ldots,L_k]{P} C'$$

to describe the process of forming a new cluster with C and P. Figure 4 gives a simple example of the recruiting operation.

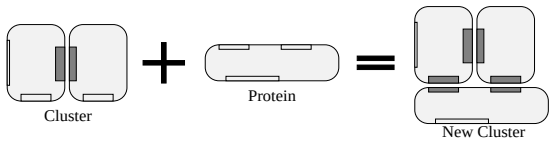


Figure 4: Recruitment

The recruitment of P by C with links $L_1, \ldots, L_k$ is *possible* if: 1) The multi-set of the first projections of $L_1, \ldots, L_k$ is included in the free domains of C, and 2) The multi-set of the second projections of $L_1, \ldots, L_k$ is included in the domains of P.

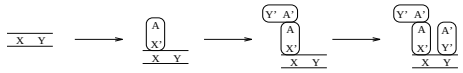
If C recruits P with links $L_1, \ldots, L_k$, the new cluster C' is obtained from C by erasing the domains of C that participate in the binding, adding those of P that do not participate in the binding, and by adding P to the members of C. Note that there is an element of non-determinism in choosing which domains of P will be bound. If additional information is available on the structures of the proteins involved, it can eventually be used to reduce the number of possible clusters.

A *medium* is a multi-set of proteins. These are the proteins available for recruitment by a cluster. Given an initial medium and a promoter I, a recruitment *chain* is an unbounded sequence: $I \longrightarrow C_1 \longrightarrow C_2 \longrightarrow C_3 \longrightarrow \cdots$ such that, at each step, either a protein is recruited by cluster C_i - and removed

from the medium - or no further recruitment is possible and the chain becomes constant (ie. $C_i + 1 = C_i$).

The set Pop_n of clusters after n recruitments is the set of n^{th} clusters from all possible recruitment chains of length n. A cluster in Pop_n contains at most n proteins, and clusters that contain less than n proteins are *stable* - they cannot grow anymore. (Note that we model *growth* of clusters, we do not, yet, account for the fact that a cluster could release a protein.)

Example 3.1 Consider the following situation. A promoter has two DNA-binding sites, named X and Y, and the medium is composed of five proteins: one copy of protein P with binding domain $\{X'\}$; one copy of protein Q with domains $\{X', A\}$; and three copies of protein R with domains $\{Y', A'\}$. Assuming that the binding relation is given by all pairs of the form (D, D'), one recruitment chain is:



Note that in the middle step, the new protein could have been recruited in two other different ways, with link (Y, Y'), or both (A, A') and (Y, Y'). ∎

3.3 Probability distribution on Pop_n

We next define a probability distribution on a population of clusters. This distribution will depend essentially on the probability of recruiting events, given a medium M:

$$p\left(C \xrightarrow[L_1,\ldots,L_k]{P} C', M\right).$$

The probability of such an event depends on many variables such as the concentration of the molecule P in the medium, the number of other binding partners in M that can compete with P, the stability of the clusters, affinities between domains, the physical realizability of links $L_1, \ldots, L_k$, etc. The simplest model, and for which probabilities can be readily computed, is based on the following:

Given a medium M, and assuming recruitment is possible, the probability that a cluster C recruits a protein $P \in M$ with links $L_1, \ldots, L_k$, is given by the quantity $n/(cN)$, where n is the total number of domains of P proteins that can bind to a free domain of C, N is the total number of domains in the medium M that can bind free domains of C, and c is the number of different ways that C can recruit P - that is, the number of different configuration of used and unused links between C and P.

Given a sequence of recruitments $s = I \longrightarrow C_1 \longrightarrow \cdots \longrightarrow C_n$, the probability $p(s)$ of the sequence is given by the product of the probabilities of each recruiting event in the sequence, with the convention that the probability of an event $C \longrightarrow C$ is equal to 1. Fixing the initial medium I, we have:

Definition 3.1 *The probability of a cluster in* Pop_n
Let $C \in Pop_n$, then the probability $p(C) = \sum_{\substack{s \text{ is a sequence of} \\ \text{length } n \text{ from } I \text{ to } C}} p(s)$. ■

Proposition 3.2 Definition 3.1 yields a probability distribution on Pop_n. ■

3.4 Random variables

It is now possible to define and compute the value of random variables defined on clusters. The most interesting are the number $N(C)$ of a specific protein P in a cluster. The mathematical expectation of these quantities in Pop_n is given by:

$$E_n(N) = \sum_{C \in Pop_n} N(C)p(C).$$

An interesting property of these quantities is expressed in the following proposition. It gives a recursive formula for the computation of $E_n(N)$. With this formula, the value of the expectation can 'accumulate' as soon as clusters are formed by an algorithm. Since a population of clusters can be very large, it might be wise to monitor the increasing approximations of $E_n(N)$ during the computation.

Proposition 3.3 $E_n(N) = E_{n-1}(N) + \sum_{\substack{s \text{ has recruitment} \\ \text{of } P \text{ as last event}}} p(s)$. ■

Example 3.2 Consider the problem described in Example 3.1, where the medium is composed of $1 \times \{X'\}, 1 \times \{X', A\}$ and $3 \times \{Y', A'\}$. The different clusters of Pop_1, Pop_2 and Pop_3 can be constructed in a tree-like fashion, with the promoter at the top (see Figure 5).

The elements of Pop_3 are found on either the third level - downward - of the tree, or at the end of branches that did not grow. The probability of the different clusters for a population can be obtained by labelling the branches of the tree with the different probabilities of recruitment.

For example, in labeling the top right arrow illustrating the recruitment of R from the initial promoter, there are $n = 3$ domains of R proteins that can bind to the promoter, $N = 5$ domains in the medium that can bind to the promoter, and $c = 1$ way that the promoter can bind an R protein, yielding a probability of $3/5$ for this particular event.

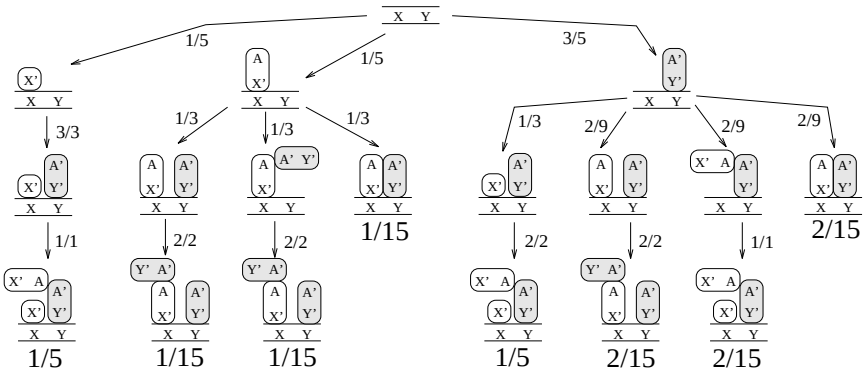


Figure 5: Construction of Pop_1, Pop_2 and Pop_3

The probability of a particular sequence of recruitment is then obtained by multiplying all the numbers along the sequence. The number under each Pop_3 cluster corresponds to the probability of obtaining this particular cluster with this particular sequence of recruitments. In the tree, we have highlighted all occurrences of protein R. The successive values of $E_n(N_R)$ are: $0, 3/5, 3/5+2/5$ and $3/5 + 2/5 + 2/15 + 2/15$, yielding an expectation of about 1.27 for the number of R proteins.

4 Virtual Experiments

4.1 The problem

Aberrant expression and structural alteration of transcription factors are frequent, primary molecular mechanisms in oncogenesis. Our target is the 11:22 chromosomal translocation found in human Ewing's tumors that results in the expression of a chimeric protein. This fusion protein combines the DNA-binding domain of the known transcriptional factor Fli-1, with domains - QSY repeats - derived from the EWS (for Ewing's) gene [12]. In vitro, the EWS-Fli-1 fusion protein has been shown to activate transcription at a rate 30 times fold the rate observed with the normal Fli-1 product [3]. Our goal is to perform experiments combining transcriptional factors with, or without, this protein.

4.2 The Virtual Laboratory

Setting up the virtual laboratory proved harder in practice than in theory. Many simplifying decision were taken with respect to biological reality in order

to track down the combinatorial complexity of the problem. The first task was to list, with the help of databases, known transcriptional factors, including two of the basal transcriptional machinery and RNA Pol II, which could be present in human hematopoietic cells [13,14](see Appendix). For each of the 32 chosen proteins, we identified known domains[5] presumably involved in protein-protein or protein-DNA interactions. These interactions are listed in Table 1.

Table 1 cannot be used immediately as a formal table of interaction: structural information on the DNA-binding sites of the chosen promoter reveals that NF-kB and Ets-1 sites are overlapping: they also must be considered of the same type, implying equality of types of both NF-kB and Ets-1 binding sites. Finally, in some cases pairs of proteins are experimentally known to interact, without the explicit knowledge of the domains through which these interactions occur. In order to take this fact into account, we added pairs of domains D? and D?', adding these to the set of domains of participants in known interactions.

As a target promoter in our virtual laboratory, we artificially introduce in blood cells a gene under the control of the Human Immunodeficiency Virus-1 (HIV-1) promoter sequence [15] - which has a complex set of protein binding sites - retaining the following multi-set of domains: { 2× AP-1, 1× TBP, 1× Myb, 1× TCF, 2× NF-kB (including 2× Ets-1), 1× USF, 3× Sp1, 1× YY1 }.

A *virtual experiment* consists in choosing the relative concentration n_P of each of the protein P in the medium, where n_P is a non-negative integer. This choice can reflect either a coarse estimate of relative concentration, or a measure of the relative affinities of different proteins for the same target. For proteins that are not currently under study, we set $n_P = 0$.

4.3 Examples of virtual experiments

In the first experiment, we want to compare the effect of binding EWS-Fli-1 or Fli-1 on the Ets Site of the promoter. The medium consists here of Pol II and the set $\{1, 6, 10, 13, 14, 15, 16\}$ where the numbers refers to proteins in Appendix. A QSY' domain was assigned to 10, 13 and 15. Results on the expected number of Pol II are shown in Table 2. These results show that the expected number of Pol II is increased at least three-fold for nearly every size of clusters, suggesting that the presence of EWS-Fli-1 could increase transcription as observed *in vitro* [3].

Experiment 2 was designed to measure the inhibitory effect of transcriptional factor. The medium consists here of Pol II, the multi-set $\{4, 5, 18, 2\times 19, 24\}$, and graded concentrations of SCL. The factor SCL was chosen for the sake of its Basic domain, theoretically competing with the RNA Pol II repeats

domain of Pol II for acidic domains. Results on the expected number of Pol II are shown in Table 3. An inhibitory dose-dependant effect is observed.

5 Conclusions

Our biological model is based on the assumption that the function of several regulatory proteins could be defined by their ability to form transient multi-molecular clusters. The formation of clusters occurs through non-covalent interactions which are dependent on the protein 3D structure. In this paper we described computational tools, based on combinatorial analysis of cluster formation, that can provide guidelines to experiments on the function of regulatory proteins in biological processes.

Although presently limited by our software knowledge on domains, we believe that this approach could be useful in the design of laboratory experiments that could tap into the tons of biological data accumulating on computing networks.

We are also aware of the elementary aspect of defining probabilities of reruitment on purely combinatorial grounds. Further implementations will incorporate more sophisticated knowledge on cluster formation, which could take into account the equilibrium constant of each two-by-two relation.

Table 1: The binding relation

Name of protein domain	Name of target interaction domain - Protein or DNA -
A. Protein-protein interactions	
1. Basic domain	Acidic domain
2. RNA Pol II repeats	Acidic domain
3. Gln-Ser-Tyr domain (QSY)	QSY'
4. Glutamine rich domain (Q-rich)	Q-rich
5. Helix-loop-helix (HLH)	HLH
6. Hydrophobic domain	Hydrophobic domain
7. Leucine zipper (Leu zip)	Leu zip
8. Proline rich domain (Pro-rich)	Pro-rich
9. Rb pocket	Rb pocket interaction
10. D1 - D6	D1' - D6' (experimental results)
B. Protein-DNA interactions	
1. AP-1 BD	AP-1 Site
2. Ets-1 BD	Ets-1 Site
3. USF BD	USF Site
4. NF-kB BD	NF-kB Site
5. Sp1 BD	Sp1 Site
6. YY1 BD	YY1 Site
7. A-T hooks (HMG)	TCF Site
8. WWW motif	Myb Site
9. TBP BD	TBP Site

Table 2: Experiment 1

n	With EWS-Fli-1		With Fli-1	
	Number of clusters	Expected number of Pol II	Number of clusters	Expected number of Pol II
3	26	0.99	9	0.00
4	92	1.33	26	0.33
5	211	1.71	41	0.48
6	335	2.30	45	0.62
7	352	2.65	38	0.77
8	257	2.98	28	0.93
9	157	3.11	22	1.05
10	118	3.19	21	1.08

Table 3: Experiment 2

n	Control		1× SCL		2× SCL		5× SCL	
	Number of clusters	Expect. number of Pol II	Number of clusters	Expect. number of Pol II	Number of clusters	Expect. number of Pol II	Number of clusters	Expect. number of Pol II
3	72	0.60	87	0.53	89	0.48	89	0.38
4	134	0.90	181	0.80	211	0.71	211	0.52
5	165	1.32	280	1.16	415	1.01	420	0.71
6	126	1.61	232	1.39	514	1.20	566	0.83
7	98	2.20	196	1.90	545	1.58	743	1.03
8	48	2.39	96	2.05	323	1.71	634	1.13
9	34	3.07	68	2.58	199	2.04	528	1.31
10	30	3.12	60	2.62	141	2.14	298	1.38

Appendix: Transcriptional Regulatory Proteins

A- Transcript. regulatory prot. which do not bind to the target DNA promoter

	SWISS-PROT NUMBER	NAME	DOMAINS
1.	P49715	C/EBP-a	Basic, Pro-rich, Leu zip
2.	Q01658	DR1	Q-rich
3.	P15923	E2A/E47	HLH
4.	?	E2f	Leu zip, Rb pocket interaction, acidic
5.	P01100	Fos	Basic, Leu zip
6.	P12980	Lyl1	HLH
7.	Q00987	MDM2	Acidic
8.	P01106	Myc	Basic, Leu zip, HLH, Q-rich, D2'
9.	P43354	NOT	Q-rich, 2× Pro-rich
10.	P04637	P53	Basic, Acidic, Pro-rich, Hydrophobic
11.	Q01094	RABP	Rb pocket interaction
12.	P06400	Rb	Rb pocket, D2, D3, Pro-rich
13.	P11831	SRF	2× Acidic
14.	P17542	SCL	Basic, HLH
15.	P21675	TAFII 250	Acidic, Pro-rich, D1'
16.	P19484	TFEB	Q-rich, HLH, Leu zip, Pro-rich
17.	P19532	TFE3	Basic, HLH, Leu zip, Pro-rich
18.	P24928	RNA Pol II	52× RNA Pol II repeats, (Catalytic domain for RNA synth.)

B- Transcript. regulatory prot. which can bind to the target DNA promoter

	SWISS-PROT NUMBER	NAME	DOMAINS
19.	P05412	AP-1	AP-1 BD, Leu zip, Q-rich
20.	P43268	E1A-F	Ets-1 BD, Q-rich, Acidic
21.	P32519	Elf-1	Ets-1 BD, Rb pocket interaction, Acidic
22.	P14921	Ets-1	Ets-1 BD, HLH, D4
23.	Q01844	Ews	31× QSY
24.	Q01543	Fli-1	Ets-1 BD, HLH
25.	P01242	Myb	WWW motif, Leu zip, Acidic
26.	P19838	NF-kB	NF-kB BD, D5'
27.	P17947	Pu.1	Ets-1 BD, HLH, 3× Acidic, Q-rich, D3', D4
28.	P08047	SP1	Sp1 BD, Q-rich, D5, D6'
29.	P20226	TBP	TBP BD, Q-rich, Basic, D1, D4'
30.	P36402	TCF-1	HMG, Pro-rich
31.	P22415	USF	USF BD, HLH, Leu zip
32.	P25490	YY1	YY1 BD, Acidic, D6
33.	-	Ews-Fli-1	Ets-1 BD, 31× QSY

References

1. Baron M., Norman D.G., Campbell L.D., Trends Biochem. Sci. **16**, 13-17 (1991).
2. Patthy L., Current Opinion in Structural Biology **1**, 351-361 (1991).
3. May W.A., Lessnick S.L., Braun B.S., Klemsz M., Lewis B.C., Lunsford L.B., Hramas R., Denny C.T., Molecular and Cellular Biology **13**, 7393-7398 (1993).
4. Fischman, J., Science **273**, 591-593 (1996).
5. Bairoch A., Bucher, P., Hofman, K., Nucleic Acids Res. **24**, 189-196 (1996).
6. Faux, M.C., Scott, J.D., Cell **85**, 9-12 (1996).
7. Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J.D., Molecular Biology of the Cell, (3rd Ed.), Garland Pub. Inc., NY, (1994).
8. Boulikas, T., Critical Reviews in Eukaryotic Gene Expression **5**, 1-77 (1995).
9. Tijian R., Maniatis T., Cell **77**, 5-8 (1994).
10. Cohen G.B., Ren R., Baltimore D., Cell **80**, 237-248 (1994).
11. Branden C., Tooze J., Introduction to Protein Structure, Garland Pub. Inc., NY, (1991).
12. Rabbits T.H., Nature **372**, 143-149 (1994).
13. Faisst S., Meyer S., Nucleic Acids Res. **20**, 3-26 (1991).
14. Wingender, E., Dietze, P., Karas, H., Knuppel, R., Nucleic Acids Res. **24**, 238-241 (1996).
15. Gaynor R., AIDS **6**, 347-363 (1992).