Review

SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways

Marko Djordjevic *

Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, United States

Abstract

Systematic Evolution of Ligands by EXponential enrichment (SELEX) is an experimental procedure that allows extraction, from an initially random pool of oligonucleotides, of the oligomers with a desired binding affinity for a given molecular target. The procedure can be used to infer the strongest binders for a given DNA or RNA binding protein, and the highest affinity binding sequences isolated through SELEX can have numerous research, diagnostic and therapeutic applications. Recently, important new modifications of the SELEX protocol have been proposed. In particular, a modification of the standard SELEX procedure allows generating a dataset from which protein–DNA interaction parameters can be determined with unprecedented accuracy. Another variant of SELEX allows investigating interactions of a protein with nucleic-acid fragments derived from the entire genome of an organism. We review here different SELEX-based methods, with particular emphasis on the experimental design and on the applications aimed at inferring protein–DNA interactions. In addition to the experimental issues, we also review relevant methods of data analysis, as well as theoretical modeling of SELEX.

Contents

1. Introduction

Systematic Evolution of Ligands by EXponential enrichment (SELEX) is a combinatorial chemistry procedure that allows rapid selection, starting from a large initial library of oligonucleotides, of the oligos that have appropriate binding affinity to a given molecular target (Tuerk and Gold, 1990).[1]

* Tel.: +1 614 292 6159; fax: +1 614 247 6643.
  E-mail address: mdjordjevic@math.ohio-state.edu.

---

[1] SELEX was also termed SAAB for selected and amplified binding sites (Blackwell and Weintraub, 1990) and CASTing for cyclic amplification and selection of targets (Wright et al., 1991), or simply *in vitro* selection (Oliphant et al., 1989; Ellington and Szostak, 1990).

Molecular targets in SELEX protocol can be either proteins or small molecules. The oligonucleotide library can consist of either single stranded oligonucleotides (RNA, ssDNA, modified RNA or modified ssDNA), or double stranded DNA (dsDNA). Little is known beforehand about the binding properties of a target molecule in many cases, so one most often starts with a large library of random oligonucleotides. Strong binders are selected from the initial library, by repeated cycles of target binding, selection and amplification.

The first SELEX experiments were performed more than fifteen years ago (Oliphant et al., 1989; Ellington and Szostak, 1990; Tuerk and Gold, 1990) and, up to now, there have been quite a few reviews that addressed SELEX experiments and their applications (e.g. Gold, 1995; Gold et al., 1997; White et al., 2000; Bowser, 2005; Bunka and Stockley, 2006). These reviews concentrated on SELEX procedure in which single stranded oligonucleotides were used, while the experiments and applications involving dsDNA were mainly not discussed. This bias is mostly due to the fact that single stranded oligos obtained through SELEX have important diagnostic and therapeutic applications. In particular, single stranded oligos that bind with strong binding affinity can be identified for a large variety of molecular targets. Those strong binders can, for example, be used as alternatives to antibodies in many applications.

On the other hand, SELEX is also a very important tool to infer interactions of proteins with dsDNA. The main reason is that one often has a protein that interacts with dsDNA *in vivo*, but whose binding specificity is unknown. SELEX was performed in many such cases in order to identify dsDNA sequences that are the strongest (consensus) binders to the protein of interest (e.g. Oliphant et al., 1989; Blackwell and Weintraub, 1990; Wright et al., 1991). Further, recently developed modifications of the standard SELEX procedure allow robust generation of a dataset, from which protein–DNA interaction parameters can be determined with high accuracy (Roulet et al., 2002; Djordjevic and Sengupta, 2006), and, consequently, binding affinity for any DNA segment can be determined. Another recently implemented SELEX-based procedure (Shtatland et al., 2000), allows experimental identification of genome derived nucleic-acid segments that are high affinity binders to a given target protein. More generally, the ability to accurately characterize interactions of nucleic acid sequences with proteins of interest is a crucial step toward understanding regulatory pathways (see e.g. Wei et al., 2004), which is, in turn, posed as one of the most important problems in molecular biology.

In this review, we will discuss different SELEX-based methods, their applications, data analysis and computational modeling. We will especially discuss appropriate experimental design, particularly in the context of different desired experimental outcomes. We will also focus on the applications of SELEX when dsDNA is used, which should complement recent reviews that address SELEX. However, we note that most of the discussion here also applies to single stranded oligonucleotides. For completeness of the discussion, we also briefly overview different aspects of SELEX that are not covered in more detail here, and point the reader to appropriate reviews.

2. SELEX protocol

The scheme of the SELEX procedure is shown in Fig. 1, and the experiments are typically performed as follows. One prepares a library of oligonucleotides that can be amplified, and the library is incubated with a target of interest. Next, the oligonucleotides that are bound by the target are separated from those that are not bound (e.g. by gel shift or filtration through nitrocellulose), which is called the selection step. Selected oligonucleotides are then amplified. The amplification is performed by PCR in the case of DNA, and by RT-PCR followed by *in vitro* transcription in the case of RNA. One cycle of target binding, selection and amplification is called a SELEX round. The SELEX rounds are repeated several times, and some of the oligonucleotides selected in the final round of the experiment are sequenced.

A large variety of molecules can be targets in SELEX experiments. When the library consists of dsDNA, targets are typically proteins that interact with DNA as a part of their natural function. In the case of single stranded oligonucleotides, a wide range of proteins and small molecules are used as targets in SELEX experiments.

The initial library of oligonucleotides typically consists of a large number (10^{15} to 10^{16}) of random sequences. In principle, larger libraries of up to 10^{20} oligonucleotides are technically feasible (Gold, 1995), but are rarely used in practice. Each oligonucleotide consists of a central region of random sequence, which is flanked by two regions of fixed sequence that enable amplification. The length of the random region is typically between 20 and 30 bp, while each flanking region is typically 15–25 bp long.

The size of the oligonucleotide library is so large, that in many cases it completely saturates the relevant sequence space. For example, each possible sequence segment of length 20 will appear, on average, about 10^4 times in the library of size 10^{15}.

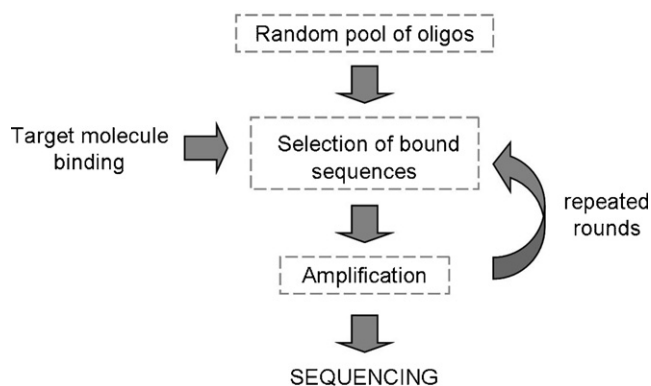Fig. 1. A scheme of the standard SELEX procedure. The starting pool of sequences consists of random oligonucleotides. The oligonucleotide pool is then gradually enriched for the high affinity binders, by repeated rounds of target molecule binding, selection and amplification. After a certain number of rounds is performed, some of the oligos selected in the last round of the experiment are sequenced.

Therefore, since binding sites of transcription factors (proteins that bind DNA and regulate gene expression) are typically less than 20 bp long, each possible sequence variant to which this protein can bind will be represented in a large number in the SELEX library. A consequence of the presence of many copy numbers of each sequence variant is that stochastic effects (e.g. loss of sequence variants due to random fluctuations) can be generally neglected in SELEX. Accordingly, stochastic effects were not taken into account in numerical simulations (Irvine et al., 1991; Vant-Hull et al., 1998) and theoretical models (Djordjevic and Sengupta, 2006) of SELEX experiments.

2.1. In vitro selection versus in vitro evolution

The term "evolution" in the name SELEX (Systematic Evolution of Ligands by EXponential enrichment) implies that both selection and mutation are important in the SELEX procedure. Since each PCR amplification necessarily introduces some level of mutation, the term "evolution" is in principle accurate. However, as we will discuss below, mutations are normally not an important effect in SELEX, so it is more useful to think about SELEX as *in vitro* selection.

To observe that the effect of mutations is small, the following estimate is useful. High fidelity DNA polymerase, which is typically used in SELEX, has a mutation rate of $\sim 10^{-4}$ per cycle per base (Eckert and Kunkel, 1990). Furthermore, let us assume that a total of seven SELEX rounds are performed, that there are 10 PCR cycles per round and that the length of DNA sequences is 25 bp. With these (typical) parameter values, a DNA sequence selected in the end of the experiment experiences, on average, a total of less than one mutation during the whole experiment (i.e. $10^{-4} \times 25 \times 10 \times 7 < 1$). Therefore, the effect of mutations in SELEX is generally small under typical experimental conditions, and quantitative models of SELEX do not take into account mutations (Irvine et al., 1991; Vant-Hull et al., 1998; Djordjevic and Sengupta, 2006).

Further, only selection without mutation is sufficient for the experiment to identify the 'solution' (i.e. the strongest affinity binders), since the initial large pool of oligonucleotides typically saturates the relevant sequence space. That is, the strongest binders already exist in the starting pool, and they do not have to be generated through mutations. The small effect of the mutations justifies the term *in vitro* selection, which is sometimes used as an alternative term for SELEX (see e.g. Oliphant et al., 1989; Ellington and Szostak, 1990; Famulok et al., 2000).

On the other hand, with respect to the importance of mutations, SELEX (or *in vitro* selection) is quite different from *in vitro* evolution experiments (Lorsch and Szostak, 1994; Dubertret et al., 2001; Peng et al., 2003). In those experiments, one purposefully starts from a (relatively) small initial pool of random sequences (e.g. 10^5 oligos), so that there is a very low probability that the strongest binders exist in this pool (Dubertret et al., 2001). Alternatively, one is using a random pool of normal size (10^{15} oligos), but the length of the binding sequences is large (e.g. 100 bp), so the relevant sequence space is sparsely populated (Lorsch and Szostak, 1994). In either case, mutations are necessary for the system to be able to find the solution, i.e. the best binders have to be generated through mutations, before they can be selected. Consequently, in *in vitro* evolution experiments, a high mutation rate is purposefully introduced.[2]

3. Selecting highest affinity binders from SELEX

In many cases the aim of *in vitro* selection is to identify the strongest (consensus) binders for a given target molecule. In this section we discuss some considerations which are relevant when SELEX is performed with that aim.

For simplicity, we here assume that the target is a protein. A protein can interact with nucleic acids either sequence-specifically, or non-specifically. Sequence-specific interactions are based on hydrogen bonds and Van-der Waals interactions, while non-specific interactions are due to electrostatic interactions alone (Jones et al., 2001; Gerland et al., 2002 and references therein; Magee and Warwicker, 2005). When the sequence of an oligonucleotide binder is far from the consensus, interaction of the protein with DNA becomes sequence-independent (Winter et al., 1981). Therefore, since a starting SELEX library consists of random sequences, most of the sequences in the starting pool will non-specifically interact with the target protein.

In general, a number of non-specific binders will also be selected in each round of the experiment, in addition to the selected sequence specific binders. The selection of non-specific binders is a consequence of two facts. First, a number of sequences will be non-specifically bound by the protein. Second, during the selection step, it is not possible to completely separate sequences that are bound by the protein, from those that are not bound. The second effect is termed background partitioning (Tuerk and Gold, 1990; Vant-Hull et al., 1998).

Non-specific interactions are typically characterized by several orders of magnitude smaller binding affinity compared to sequence specific interactions (Stormo and Fields, 1998). Also, background partitioning probability, i.e. the probability to select a sequence that is not bound by the protein, is likely low, e.g. 10^{-3} (Tuerk and Gold, 1990). However, these small numbers do not imply that the presence of non-specific binders can be neglected. For example, about 10^{12} non-specific binders will be selected in the first round of the experiment, just due to background partitioning, assuming the background partitioning probability of 10^{-3} and a starting library size of 10^{15} sequences. On the other hand, with protein–nucleic acid ratio of 10^{-3}, which is typical for SELEX experiments (e.g. Tuerk and Gold, 1990), less than 10^{12} specific binding sequences will be selected. Consequently, after the first round of the experiment, the number of non-specific binders is typically comparable or even larger than the number of specific binders. As a practical consequence, one cannot perform just one round of SELEX

---

[2] This is most often done by mutagenic PCR (Bartel and Szostak, 1993) and/ or by performing a large number of PCR rounds.

with an aim of obtaining a pool of weaker, but still sequence specific binders to a protein of interest. That is, multiple rounds of selection are generally needed in order to eliminate the 'noise' due to non-specific binding.

Another important issue is how the affinity distribution of the selected sequence-specific binders changes through different experimental rounds. While it is intuitively evident that the affinity generally increases with more performed rounds, quantitative behavior depends on both experimental parameters and properties of protein–nucleic acid interactions. In a simple case, which is realized when the protein to nucleic acid ratio is sufficiently small, the average binding affinity of the selected sequence specific binders increases exponentially with the number of performed SELEX rounds (Schneider et al., 1993; Vant-Hull et al., 1998; Djordjevic and Sengupta, 2006). Such exponential increase in binding affinity, during the first few rounds of the experiment, justifies the term "exponential" in the name of Systematic Evolution of Ligands by EXponential enrichment. Finally, after a certain number of rounds are performed, the maximum of the affinity distribution of the selected binders reaches an upper limit, which is determined by the affinity of the strongest binder in the starting random library. At that point, most of the sequences in the selected pool will consist of the strongest binders.

A practically important question is how many SELEX rounds have to be performed in a given experiment. In order to select the strongest binders from the initial pool, non-specific binders have to be eliminated, and affinity of sequence-specific binders has to reach the upper limit discussed above. In principle, the change of affinity distribution, including the number of non-specific binders, can be calculated from the computational models (Vant-Hull et al., 1998; Djordjevic and Sengupta, 2006). The problem is, however, that the parameters of protein–nucleic acid interactions are *a priori* unknown in practice, so one cannot use such a calculation to decide when the experiment should be stopped. Due to that, the best criterion for the number of rounds needed to be performed is a binding analysis of selected oligos against the protein target, through different rounds of SELEX (see e.g. Schneider et al., 1993). An (approximate) saturation of the measured binding affinity is an indication that the experiment can be stopped. In practice, between 5 and 15 SELEX rounds are typically performed.

Finally, if the purpose is to select the strongest binders from the initial random pool, it is of course necessary that these sequences are not eliminated in some of the selection steps. Since the fraction of strong binders increases with the number of performed rounds, the 'risk' of eliminating them is greatest in the early rounds of the procedure. This is why in practice one often starts with a higher protein to oligonucleotide ratio at the beginning of the experiment, but then gradually decreases this ratio as the number of performed rounds increases (e.g. He et al., 1996). The reason behind such a procedure is to decrease the stringency of selection in the early rounds of the experiment, but to increase it later in the experiment, when the strongest binders are present in larger quantity. Such a procedure was also suggested by numerical simulations (Irvine et al., 1991).

3.1. Highest affinity single stranded oligonucleotides

When single stranded oligonucleotides are used in the SELEX procedure, high affinity binders for a large variety of target molecules can be obtained. Specifically, single stranded oligonucleotides that bind with high affinity to diverse targets such as organic dyes, amino acids, antibiotics, peptides, proteins or vitamins have been obtained by SELEX experiments (see Tombelli et al., 2005 and references therein). The ability to isolate a strongly binding oligo for almost any molecular target is due to the ability of single stranded oligonucleotides to incorporate small molecules into their structure, or to integrate into the structure of larger molecules (Hermann and Patel, 2000). Single stranded oligonucleotides obtained as the strongest binders in SELEX are often called aptamers (derived from the Latin "aptus" meaning "to fit").

Aptamers have important research, diagnostic and therapeutic applications. The efficiency of obtaining aptamers for various applications has been recently enhanced, by automating the SELEX protocol (Cox and Ellington, 2001). Applications of aptamers were extensively discussed in several recent reviews (e.g. Kopylov and Spiridonova, 2000; Bowser, 2005; Bunka and Stockley, 2006), so we will not consider them in detail here. However, for completeness, we briefly overview these applications below, and point the interested reader to the relevant references.

In research applications, SELEX is widely used to recognize RNA motifs to which a given protein strongly binds (e.g. Shi et al., 1997). If the protein of interest binds RNA under natural conditions, the knowledge of such a motif may elucidate functional roles of this protein, for example in a given regulatory pathway. Further, SELEX-like methodology can also be used to identify oligonucleotides with a desired enzymatic activity. One should note that catalytic RNA structures are often significantly more complex than simple aptamers. Consequently, mutagenic PCR (i.e. *in vitro* evolution) is often used in the final rounds of such experiments to enhance the finding of an optimal catalyst. Catalytic nucleic acid aptamers have been reviewed in detail in Wilson and Szostak (1999) and Gilbert and Batey (2005).

Further, aptamers are often able to interfere with a biological function of the target molecule, since a high affinity interaction of an oligo with a protein target may overlap with an active site of the protein. If such aptamers are expressed in living cells, they can be used to investigate intracellular signal transduction pathways. Recently developed techniques that facilitate the intracellular applications of aptamers are reviewed in Famulok et al. (2000) and Ulrich (2005).

Aptamers can also be used in diagnostic applications, as alternatives to antibodies, which is based on their ability to bind target molecules with high affinity and specificity. Diagnostic applications of aptamers have been reviewed in Jayasena (1999) and Brody and Gold (2000). Related to the diagnostic applications, aptamers can be used in analytical chemistry, with applications that range from separation techniques to biosensors. The analytical applications of aptamers have been reviewed in Tombelli et al. (2005) and Ravelet et al. (2006).

In therapeutic applications, several aptamers, which are known to inhibit various target proteins, are currently tested in pre-clinical and clinical trials. Various therapeutic applications of aptamers have been reviewed in White et al. (2000), Ulrich et al. (2006) and Goringer et al. (2006). Finally, efficiency of aptamers in diagnostic and therapeutic applications can be enhanced by chemical modifications that provide resistance of oligonucleotides against enzymatic degradation in body fluids, which has been reviewed in Kusser (2000) and Kainz et al. (2006).

3.2. Highest affinity dsDNA sequences

In addition to isolation of aptamers, SELEX was used in many cases (e.g. Oliphant et al., 1989; Blackwell and Weintraub, 1990; Wright et al., 1991; Robison et al., 1998 and references therein; Beinoraviciute-Kellner et al., 2005; Yagura and Itoh, 2006) in order to identify the consensus (strongest) binding sequence to a given dsDNA binding protein. In most cases, the dsDNA binding protein of interest binds to genomic DNA and regulates expression of genes. Consequently, the knowledge of the consensus sequence is usually further used in an attempt to infer likely protein binding sites in the genome. Genomic targets of the protein may, in turn, point to possible involvement of this protein in different regulatory pathways.

However, the knowledge of the consensus sequence is usually not sufficient to determine the binding sites of protein in the genome. The main reason is that binding sites in genomic DNA typically show significant sequence variations (Stormo, 2000). The biological reason for this sequence variability is that different binding site affinities can lead to different (desired) levels of gene expression. Moreover, it is useful to observe that a SELEX library is typically much larger than the size of a genome, so the consensus sequence may not be present in the genome at all. Therefore, a direct match with the consensus sequence cannot be used to identify binding sites in most cases. As an alternative, one may attempt to identify binding sites by allowing certain number of mismatches to the consensus sequence. However, a general problem with this approach is that different positions in a binding site, as well as different mismatches at a given position, can contribute very differently to protein–DNA interaction energy (Stormo and Fields, 1998; Fields et al., 1997). That is, while a certain mismatch can almost completely abolish sequence-specific binding, another mismatch may change the binding energy by only a small amount. Therefore, one needs a more complete set of protein–DNA interaction parameters, in order to accurately predict binding affinity of a protein to dsDNA segments.

The interaction of proteins with dsDNA can be quantified by using the so-called independent nucleotide approximation (Stormo and Fields, 1998). In this approximation, the binding energy of a protein to a dsDNA sequence is equal to the sum of contributions due to the presence of a given base at a given position in the binding site. Although there are some examples where binding at certain positions shows dependence on dinucleotide pairs (Man and Stormo, 2001; Bulyk et al., 2002; O'Flanagan et al., 2005), the independent nucleotide approx-

imation provides a very good parameterization of binding energy in most cases (Takeda et al., 1989; Sarai and Takeda, 1989; Benos et al., 2002). In order to measure relative binding affinities of proteins to dsDNA sequences within the independent nucleotide approximation, one needs a total of $3L$ independent parameters (L is the binding site length).[3] Importantly, the interaction parameters can be accurately determined from a modified version of the SELEX experiment, and we will discuss this in detail in the following sections.

Finally, it is important to note that interactions of proteins with single-stranded oligonucleotides are, in principle, more complex than protein–dsDNA interactions. This is due to the fact that interactions of proteins with single-stranded oligos generally significantly depend on RNA secondary and tertiary structures (Jones et al., 2001). In particular, proteins tend to interact with RNA in the locations where RNA forms complex secondary structure elements such as stem-loops and bulges (Nagai, 1996). Consequently, the independent nucleotide approximation is generally not appropriate for interactions of proteins with single stranded oligos, and protein–single stranded oligonucleotide interactions are not straightforward to parameterize. Due to this, we will in the next two sections concentrate on inferring protein–dsDNA interaction parameters from SELEX-based experiments. We will, however, return to protein–RNA interactions in the section on genomic SELEX.

4. Problems in determining protein–DNA interaction parameters from standard SELEX procedure

As we discussed in the previous section, one needs a set of protein–DNA interaction parameters to be able to accurately quantify binding specificity of a protein to dsDNA. Within the independent nucleotide approximation, these parameters can be written in a form of a matrix with dimension $4L$, which is called weight matrix (Stormo and Fields, 1998; Stormo, 2000). Individual weight matrix elements are proportional to the contribution to the binding energy due to the presence of a certain base at a certain position in a binding site (Stormo and Fields, 1998; Djordjevic et al., 2003). Therefore, in the case of interactions of dsDNA with proteins, an important goal is to accurately determine weight matrix elements.

One possibility for determining the weight matrix is from a set of aligned binding sites assembled in biological databases (Wingender et al., 2001; Salgado et al., 2004). However, the majority of the weight matrices determined in this way provide a low level of both specificity and sensitivity (Frech et al., 1997). In particular, there tends to be a large number of false positives in searches using most of the weight matrices (Robison et al., 1998; Stormo, 2000). This problem is typically attributed to a non-suitable dataset from which most weight matrices are constructed (Frech et al., 1997) because, first, for most DNA binding proteins, only a few binding sites are available in databases (Wingender et al., 2001; Davuluri et al.,

---

[3] One should observe that there is one parameter for each possible mismatch from a reference sequence.

2003; Salgado et al., 2004), which is insufficient to accurately determine protein–DNA interaction parameters (O'Flanagan et al., 2005). Second, binding sites from databases are often assembled under diverse and ill-characterized conditions (Djordjevic et al., 2003). Therefore, it is highly desirable to be able to generate a better dataset in order to improve the accuracy of weight matrices.

As an alternative to using binding sites assembled in biological databases, it appears that SELEX experiments can be suitable for generating an appropriate dataset under controlled (uniform) conditions. As we discussed above, the standard SELEX procedure can be used in order to efficiently infer the strongest binder. However, can this procedure also be used to generate a suitable dataset from which an accurate weight matrix can be determined?

Actually, it appears that the standard SELEX procedure often fails in practice, in terms of providing a dataset for accurate weight matrix construction. For example, in a SELEX experiment performed with a bacterial transcription factor LRP (Cui et al., 1995), about 50 binding sites extracted from the last round of the experiment were used to construct a weight matrix. The weight matrix scores for each sequence were then calculated, and binding dissociation constants for each of those sequences were also experimentally measured. Weight matrix scores and dissociation constants should be directly related, since the dissociation constant is proportional to the exponent of binding energy, and the binding energy should be, in turn, proportional to the weight matrix score. However, as noted in the paper, the correlation between the dissociation constants and the weight matrix scores was quite poor. Similarly, it was noted (Liu and Stormo, 2005) that a weight matrix constructed directly from sequences extracted in a standard SELEX procedure was not able to provide a good prediction of measured binding affinities.

Additionally, a comprehensive comparison between the weight matrices from eight available SELEX experiments with *E. coli* transcription factors, and the corresponding weight matrices constructed from natural binding sites was reported (Robison et al., 1998). In this comparison, large discrepancies between the weight matrices derived from natural binding sites and those derived from SELEX were reported in seven out of those eight cases. Therefore, it appears that the inference of accurate interaction parameters from the standard SELEX procedure is more an exception than a rule.

Why does the standard SELEX procedure appear to fail in so many cases? Actually, as we will discuss below, there is a systematic problem with the standard SELEX experiment in terms of generating a dataset suitable for inferring accurate protein–DNA interaction parameters. To understand this, it is useful to consider the kind of dataset needed to construct an accurate weight matrix. First, the 'noise' in the dataset has to be low, so a successful experiment has to eliminate non-specific binders from the dataset. Second, over-selection must not happen, i.e. the selected sequences should not consist of only the strongest binding sites. To understand the second point, it is useful to take the limit in which the dataset consists from only the consensus binder, when it is evident that the weight matrix

elements cannot be obtained from such information. A more detailed statistical analysis also shows that a significant fraction of medium affinity and weaker affinity binding sequences are needed for an accurate determination of weight matrix elements (Roulet et al., 2002).

Actually, the two requirements stated above, i.e. the elimination of non-specific binders and the absence of over-selection, are hard to reconcile within the standard SELEX procedure. This is a consequence of the fact that the selected sequence-specific binders rapidly reach the highest affinity binding sites, and non-specific binders may not be eliminated from the pool of selected sequences by that time. A quantitative model of the standard SELEX procedure actually shows that there exists a range of realistic parameters for which either of the two problems exists in every SELEX round (Djordjevic and Sengupta, 2006). Even for the parameter values for which this is not the case, it is very difficult to reliably predict when to stop the experiment in practice, i.e. to determine in which SELEX round the noise is eliminated while the over-selection has not happened yet. This is due to the fact that in most cases one uses a target protein for which protein–DNA interaction parameters are not known *a priori*, so the appropriate number of rounds cannot be simply calculated from a quantitative model of the experiment. In the next section, we will review how the SELEX procedure can be appropriately modified in order to allow a robust generation of a dataset from which accurate protein–DNA interaction parameters can be determined.

5. Fixed stringency/high-throughput SELEX experiment

To understand how to appropriately modify SELEX, we will first discuss the selection of oligos. Probability that a sequence S is bound by the protein, and consequently selected in the next round of the experiment, is given by the expression $[c]/(K_d(S) + [c])$ (Djordjevic et al., 2003). Here $[c]$ and $K_d(S)$ are the concentration of free protein and the binding dissociation constant of the sequence S, respectively. Therefore, the selection stringency is determined by the concentration of free protein in solution.

Most SELEX experiments are performed so that the total amount of protein and DNA is the same in each experimental round. Since the average binding affinity of the selected sequences increases with the number of performed rounds, the amount of bound protein will increase, and consequently, the amount of free protein will decrease. This decrease of the free protein concentration leads to an increase of the selection stringency through the experiment.[4] Consequently, we will further call the standard SELEX protocol high stringency SELEX, and as we discussed in the previous sections, such a procedure is not suitable to accurately determine protein–DNA interaction parameters.

---

[4] In practice, some SELEX experiments are performed so that the total amount of protein decreases from one round to the next, and the increase of stringency is even higher in such case.

Let us now assume that instead of decreasing, the amount of free protein is constant in each round of SELEX. Since the selection stringency for any given sequence is then constant, we will further call this procedure fixed stringency SELEX. The change of the energy distribution of selected DNA sequences, for fixed stringency SELEX, can be calculated from a quantitative model of SELEX (Djordjevic and Sengupta, 2006), and is shown in Fig. 2A. The left-most point on the horizontal axis corresponds to the energy of the strongest binder, and the value of chemical potential is also indicated in the figure. The chemical potential is proportional to the logarithm of free protein concentration (Sengupta et al., 2002), and is therefore also fixed through the experiment. One can
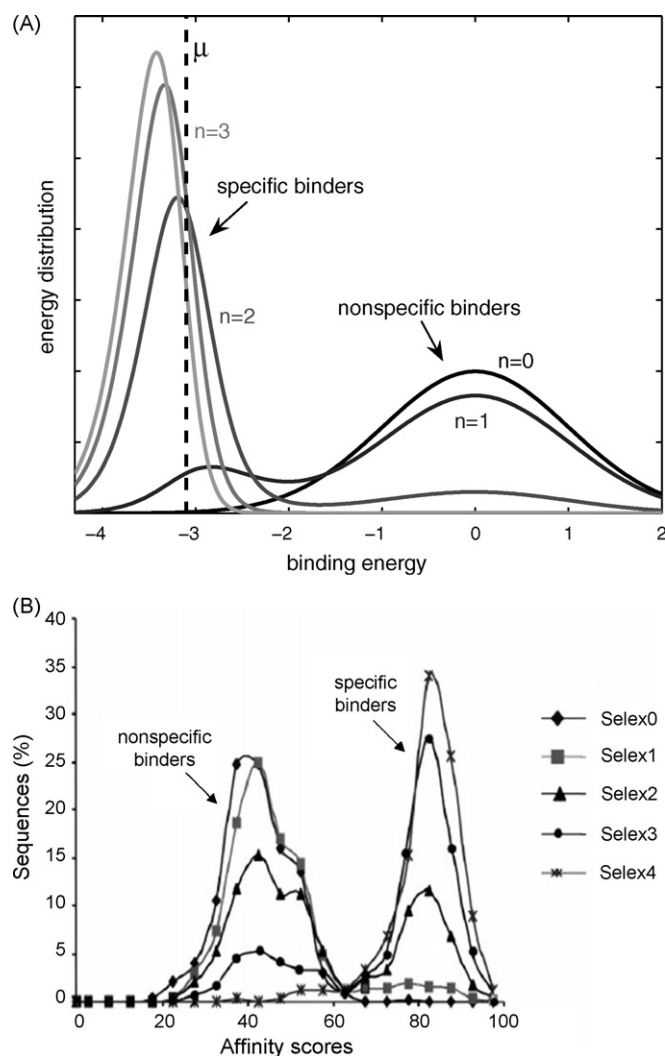

Fig. 2. Change of energy distribution through different SELEX rounds, if the amount of free protein is fixed. (A) Theoretically predicted change of energy distribution. The value of chemical potential (μ) is indicated by the vertical dashed line, and the leftmost point on the horizontal axis corresponds to the energy of the strongest binder in the initial random pool. The figure is adapted from Djordjevic and Sengupta (2006). (B) Experimentally inferred change of energy distribution. SELEX0 in the figure legend denotes the starting (random) SELEX library, while SELEX1–SELEX4 indicate, respectively, selected pools of DNA sequences after rounds one to four of the experiment. Note that the affinity scores, shown on the horizontal axis, are proportional to the negative value of the binding energy. The figure is adapted from Roulet et al. (2002).

observe from the figure that the maximum of the energy distribution for selected sequence specific binders remains in the vicinity of the chemical potential, i.e. the maximum drifts very slowly toward the higher binding energies with the additional number of performed SELEX rounds. This is in sharp contrast to the standard SELEX procedure, where the maximum of the energy distribution rapidly reaches the strongest affinity binders (Djordjevic and Sengupta, 2006), which leads to over-selection. On the other hand, one can notice that the number of non-specific binders keeps decreasing with the increase in the number of performed SELEX rounds.

The important practical implication is that in the fixed stringency SELEX, one can perform more SELEX rounds, thus ensuring that random binders are eliminated, without the risk that only the strongest sequences will be selected. It can be mathematically shown that the fixed stringency SELEX procedure leads to this desired behavior for all parameter values (Djordjevic and Sengupta, 2006). Since the procedure tolerates the large range of performed experimental rounds (in the example in Fig. 2A, any round larger than two is suitable), it is appropriate to say that this procedure is robust. Therefore, in conclusion, a SELEX experiment in which the amount of free protein is fixed through different experimental rounds, allows robust generation of a suitable dataset for accurate determination of protein–DNA interaction parameters.

How can one experimentally implement the constraint of fixed free protein amount? An answer is given by the recent experiment by Roulet et al. (2002), where the standard SELEX procedure was modified by inclusion of the radio-labeled sequence (probe) S* of moderate binding affinity. The concentration of total DNA, added to the reaction mixture as a competitor to the radio-labeled probe, was adjusted in each round of the experiment, so that a fixed fraction of the probe is bound by protein in each SELEX round. Note that radio-labeling of the probe allows one to determine the fraction of the probe that is bound by the protein. The constraint that the fraction of the bound probe is constant leads to $[c]/([c] + K_d(S^*)) = \text{const}$, where $K_d(S^*)$ is the dissociation constant of the probe. Therefore, the free protein amount ([c]) has to be constant as well, since of course $K_d(S^*)$ does not change.[5]

In the experiment by Roulet et al. (2002), a weight matrix for the target protein (a mammalian transcription factor CTF/NFI) was used to score the oligos that were extracted and sequenced in different experimental rounds. The resulting, experimentally inferred, change of energy distribution is shown in Fig. 2B. This figure shows similar behavior as the one theoretically predicted (Fig. 2A). That is, the maximum of energy distribution for selected sequence specific binders moves very slowly toward the higher binding affinities, while the number of non-specific binders keeps decreasing with the increase in the number of performed rounds. This further confirms that the fixed

---

[5] An alternative way to implement a fixed free protein amount is to keep the amount of total DNA constant, but to adjust the total protein amount so that the fraction of the bound probe is again constant in each round of the experiment.

stringency SELEX procedure can be used to reliably generate a suitable dataset.

The experiment by Roulet et al. introduced another important modification, which was to combine the SELEX procedure with the SAGE (Serial Analysis of Gene Expression) protocol (Velculescu et al., 1995). Specifically, a part of the SAGE protocol was used to link together oligos extracted from SELEX in longer DNA molecules, which can be efficiently sequenced. This technique allows one to efficiently sequence up to several thousand binding sequences (Roulet et al., 2002). The procedure was termed high throughput SELEX, or alternatively, SELEX-SAGE protocol. Such a large dataset provides an obvious advantage for a precise estimation of protein–DNA interaction parameters. Actually, with so large a dataset, one can accurately determine interaction parameters even beyond the single nucleotide approximation (O'Flanagan et al., 2005), for example contributions to the binding energy of all nearest neighbor dinucleotide pairs can be estimated. Therefore, the combination of the fixed stringency procedure with the SELEX-SAGE protocol, which we call fixed stringency/high throughput SELEX, allows both robust and accurate determination of protein–DNA interaction parameters. A database called HTPSELEX, specifically developed for storing large datasets obtained from high-throughput SELEX experiments, has recently become available (Jagannathan et al., 2006). This complements SELEX_DB (Ponomarenko et al., 2000) and TRANSFAC (Wingender et al., 2001) databases, which have been assembling the data obtained from standard SELEX experiments.

We finally discuss analysis of SELEX experimental data. It is important to note that the length of the randomized part of DNA sequences is usually larger than the length of a protein binding site. Consequently, one first has to perform a multiple local sequence alignment of the selected sequences, in order to locate statistically overrepresented motifs of certain length. The algorithms for identification of statistically overrepresented motifs are typically based on either the Gibbs search (Lawrence et al., 1993), or expectation maximization (Bailey and Elkan, 1994). The set of aligned binding sites obtained through a multiple local sequence alignment is, in a typical data analysis, used to construct an information-theory based weight matrix (Stormo, 2000). In the information-theory based method, the weight matrix elements are equal to the logarithm of the ratio of probability to observe a given base at a given position in a collection of binding sites, compared to the base background probability.

However, the information-theory weight matrix method is not appropriate to use, since it does not properly incorporate saturation in the binding probability (Sengupta et al., 2002; Djordjevic et al., 2003). That is, the information theory based method assumes that the probability that sequence S is bound by protein is given by $[c]/K_d(S)$, while the correct binding probability is given by a function with sigmoid form $[c]/([c] + K_d(S))$ (Djordjevic et al., 2003). A method which correctly incorporates saturation in binding probability, and which allows construction of an accurate weight matrix from the data generated in fixed stringency SELEX, was developed in Djordjevic and Sengupta (2006). The method was shown to lead to a significantly better false positive/false negative trade-off, as compared to the information theory weight matrix. Finally, an accurate weight matrix, which is obtained through an appropriate analysis of fixed stringency/high throughput SELEX data, can be used in order to identify putative binding sites of protein in the genome.

6. Genomic SELEX

In the previous section we reviewed a modification of SELEX that allows one to accurately determine protein–DNA interaction parameters, and to consequently predict genomic targets of a protein of interest. In this section, we will discuss a SELEX based method that allows one to directly identify genomic DNA or RNA sequences that bind with high binding affinity to a given target protein. The method was termed genomic SELEX, and similarly to the standard SELEX procedure, it is based on iterative rounds of binding, selection and amplification (Shtatland et al., 2000). However, the difference between genomic SELEX and standard SELEX is that in the former procedure the starting library is derived from the genome of the organism of interest, while in the latter procedure the library consists of random DNA oligos. A genomic SELEX library consists ideally of overlapping DNA fragments that start at all positions within the genome. The sequences in a genomic SELEX library contain fixed flanking regions that allow amplification and *in vitro* transcription, so that, if desired, the library can be expressed as RNA. Genomic SELEX libraries have been constructed for bacteriophage MS2, *Eserichia coli*, *Sinorhizobium meliloti*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and human genomic DNA (Wen and Gray, 2004; Singer et al., 1997; Ferrieres et al., 2004; Kim et al., 2003). In Singer et al. (1997) the qualities of the starting genomic SELEX libraries for *E. coli*, *S. cerevisiae* and human genomic DNA were systematically tested, and the libraries were found to contain overlapping genomic fragments that start at most of the positions in the genome.

The first genomic SELEX experiment was performed with a library derived from the *E. coli* genome, and the target protein was a known transcription factor MetJ, which binds dsDNA (Gold et al., 1997). The experiment was reported to isolate many of the binding sites known from previous genetic and biochemical studies, together with some biologically plausible, but previously unreported sites. Later, dsDNA genomic SELEX experiments were performed with the *E. coli* Cra (FruR) transcription factor (Shimada et al., 2005), and with the *S. meliloti* transcription factor FixJ (Ferrieres et al., 2004). In the experiment with Cra, all six known promoters that are repressed by Cra were identified, but none of the sequences corresponding to the known activation-type promoters were isolated.

The first RNA genomic SELEX used the bacteriophage MS2 coat protein as the target, and the SELEX was performed with a library consisting of *E. coli* DNA that was transcribed *in vitro* (Shtatland et al., 2000). It was found that the MS2 coat protein binds several *E. coli* mRNA fragments more strongly than the natural, well studied, phage mRNA site. Later, RNA genomic

SELEX was used in order to identify genomic RNA fragments that bind *Drosophila* pre-mRNA splicing factor B52 (Srp55) (Kim et al., 2003). This experiment narrowed the set of candidate B52 target genes from about 13,000 genes predicted in *Drosophila* to less than hundred genes, and some of the identified genes showed splicing defects in the B52 null mutant. While these experiments were applied to select mRNA binding partners of proteins, the genomic SELEX can also be used to experimentally detect novel non-protein–coding RNAs (ncRNA), which is reviewed in more detail in (Huttenhofer and Vogel, 2006).

While the previous examples show that genomic SELEX was successful in identifying some of the high affinity binders in the genome, there are also technical problems with this procedure. First, in the case of RNA, a problem is that the fixed flanking sequences of the library may form base pairs with the central genomic-derived fragments. This results in the formation of structures that can be selected as sites for target binding, but which do not correspond to the sequences that naturally occur in the genome, and the isolation of such artifacts was noted in (Shtatland et al., 2000). A method termed primer-free SELEX was developed to address this problem, where primer-annealing sequences are removed from the genomic library before selection, and are then regenerated to allow amplification of the selected RNA (Wen and Gray, 2004).[6]

Second, and likely a more serious systematic problem, is the possible over-selection in genomic SELEX experiments. This problem is similar to the one already discussed in standard (high stringency) SELEX. Over-selection presents a significant problem, since a number of medium or weaker affinity binding sites, which may be functionally very important, would be missed. The problem of over-selection in genomic SELEX was recognized in practice. For example, the binding affinity of the Cra transcription factor to repression type promoters is higher than the binding affinity to activation type promoters. The fact that the genomic SELEX successfully isolated known repression-type binding sequences, but none of the known activation-type binding sites (Shimada et al., 2005), clearly indicates that an over-selection occurred. Similarly, Kim et al. (2003) recognized the problem of over-selection and noted that they used fewer rounds of selection than a conventional SELEX to avoid the enrichment of just a few high-affinity winners. However, as we discussed in the case of standard (high stringency) SELEX, this generally does not solve the problem, since if too few SELEX rounds are performed, too large a number of non-specific binders will exist in the selected pool.

How can the above systematic problem in the genomic SELEX be solved? A possible solution, following from our previous discussion, is to combine genomic SELEX with the fixed stringency SELEX procedure. A labeled binding probe can be used in order to impose the fixed stringency of selection, i.e. to insure that the amount of free protein remains constant through different SELEX rounds. For example, in the case of the genomic SELEX experiment with Cro, mentioned above, some of the known weaker affinity activation-type Cro binding sites could be used as a probe. It is likely that such fixed stringency genomic SELEX experiment would recover the known activation-type promoters, which were missed by the standard genomic SELEX. It is also likely that the fixed stringency genomic SELEX would lead to a significant number of previously undiscovered, weaker – but biologically functional – Cro binding sites. Further, genomic SELEX could be combined with the high-throughput SELEX procedure, which would allow efficient isolation of most of the biologically functional binding sites even for pleiotropic regulators. For example, a pleiotropic transcription factor CAP (CRP) of *E. coli* is considered to have about 500 functional targets in the genome (Robison et al., 1998), and this, or even a ten times larger number of sequences, can be efficiently sequenced by the SELEX-SAGE protocol.

There is also a general issue relevant for the interpretation of genomic SELEX results. Genomic SELEX is able to identify a set of binding sites in a genome, or in RNA derived from a genome, that interact with a given target protein *in vitro*. However, the exact *in vivo* conditions under which such an interaction happens – if at all – still have to be determined. For example, in the case of RNA, the initial genomic SELEX library consists of RNA fragments derived from the entire genome, so the sequences selected in genomic SELEX may not be expressed at all under natural conditions. In the case of DNA, binding of proteins to the genome may depend on additional factors, such as the chromatin state or interactions with other DNA binding proteins, which is not accounted in the context of the genomic SELEX experiment. The method described in the previous section, i.e. the detection of protein binding sites through an accurate determination of protein–DNA interaction parameters from fixed stringency/high-throughput SELEX data, has the same principal limitation.

Finally, we compare genomic SELEX with a recently developed technique of Chromatin imunoprecipitation (ChIP) followed by hybridization to whole genome microarrays (ChIP-chip experiments). The method allows determining *in vivo* genomic binding sites for a given DNA binding protein (Buck and Lieb, 2004). While the method works well in some cases (e.g. Roberts et al., 2003), a significant number of false negatives is reported in some other cases (e.g. Yeung et al., 2004). A ChIP-chip experiment is performed under certain conditions, and one should keep in mind that the protein may not bind to all of its genomic binding sites under the conditions in which the experiment is done. Therefore (appropriately modified) genomic SELEX and ChIP-chip are largely complementary to each other, since the former should, in principle, identify all genomic segments with which the protein can interact, while the latter should detect binding sites that are active under the given conditions. More generally, strengths of different methods often complement each other, so genomic SELEX is best used in combination with other methods that study protein–nucleic acid interactions, such as ChIP-chip and protein binding microarrays (Bulyk et al., 2001).

---

[6] An earlier study (Vater et al., 2003) showed that the primer-annealing sequences on the 5′ and 3′ ends can be trimmed to only six and four nucleotides, respectively. However, non-naturally occurring sequences are still present in the library with this approach.

7. Conclusion

SELEX is emerging as a reliable method for inferring and quantifying protein–nucleic acid interactions. The overwhelming experimental and theoretical evidence shows that SELEX can efficiently select the strongest binders from large nucleic acid libraries, which in most cases saturate the relevant sequence space. The selection process has been well understood, both experimentally and theoretically. The strongest affinity binders obtained through SELEX have important research, diagnostic and therapeutic applications. However, the standard SELEX procedure is generally not suitable for correctly determining protein–nucleic acid binding parameters, from which the binding affinity of any sequence specific binder can be determined.

On the other hand, recent advances in the SELEX protocol allow determination of protein–DNA interaction parameters with unprecedented accuracy. In particular, fixed stringency/high-throughput SELEX procedure allows robust sequencing of a large number of medium to lower affinity sequence specific binders, under controlled conditions. The interaction parameters can be determined with high accuracy from such a dataset, by using a quantitative physical understanding of SELEX experiments, together with the developed bioinformatics methods. The determined parameters, in turn, allow the reliable detection of putative protein binding sites in genomic DNA. Therefore, such methodology can be applied to a large number of different DNA binding proteins, which would facilitate comprehensive understanding of gene regulation.

As an alternative approach, genomic SELEX can be used to determine genome-derived nucleic acid sequences that have high binding affinity to a given protein target. Genomic SELEX has been successfully used to isolate strong genomic binding sequences in a number of organisms ranging from a bacteriophage to a fly, but the method is generally biased against weaker interactions that may be functionally highly important. While not yet realized in practice, the combination of genomic SELEX with the fixed stringency/high-throughput SELEX procedure should alleviate this bias.

In summary, different methodologies based on SELEX have a large potential to reliably infer nucleic acid binding specificity of proteins and other molecular targets. This potential will likely result in the discovery of a number of novel biological regulatory loops in the future.

Acknowledgements

References

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California, pp. 28–36.

Bartel, D.P., Szostak, J.W., 1993. Isolation of new ribozymes from a large pool of random sequences. Science 261 (5127), 1411–1418.

Beinoraviciute-Kellner, R., Lipps, G., Krauss, G., 2005. *In vitro* selection of DNA binding sites for ABF1 protein from *Saccharomyces cerevisiae*. FEBS Lett. 579 (20), 4535–4540.

Benos, P.V., Bulyk, M.L., Stormo, G.D., 2002. Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Res. 30 (20), 4442–4451.

Blackwell, T.K., Weintraub, H., 1990. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science 250 (4984), 1104–1110.

Bowser, M.T., 2005. SELEX: just another separation? Analyst 130 (2), 128–130.

Brody, E.N., Gold, L., 2000. Aptamers as therapeutic and diagnostic agents. J. Biotechnol. 74 (1), 5–13.

Buck, M.J., Lieb, J.D., 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83 (3), 349–360.

Bulyk, M.L., Huang, X., Choo, Y., Church, G.M., 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc. Natl. Acad. Sci. USA 98 (13), 7158–7163.

Bulyk, M.L., Johnson, P.L., Church, G.M., 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. 30 (5), 1255–1261.

Bunka, D.H., Stockley, P.G., 2006. Aptamers come of age—at last. Nat. Rev. Microbiol. 4 (8), 588–596.

Cui, Y., Wang, Q., Stormo, G.D., Calvo, J.M., 1995. A consensus sequence for binding of Lrp to DNA. J. Bacteriol. 177 (17), 4872–4880.

Cox, J.C., Ellington, A.D., 2001. Automated selection of anti-protein aptamers. Bioorg. Med. Chem. 9 (10), 2525–2531.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., Grotewold, E., 2003. AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. BMC Bioinform. 4 (1), 25.

Djordjevic, M., Sengupta, A.M., 2006. Quantitative modeling and data analysis of SELEX experiments. Phys. Biol. 3, 13–28.

Djordjevic, M., Sengupta, A.M., Shraiman, B.I., 2003. A biophysical approach to transcription factor binding site discovery. Genome Res. 13 (11), 2381–2390.

Dubertret, B., Liu, S., Ouyang, Q., Libchaber, A., 2001. Dynamics of DNA–protein interaction deduced from *in vitro* DNA evolution. Phys. Rev. Lett. 86 (26), 6022–6025.

Eckert, K.A., Kunkel, T.A., 1990. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. Nucleic Acids Res. 18 (13), 3739–3744.

Ellington, A.D., Szostak, J.W., 1990. In vitro selection of RNA molecules that bind specific ligands. Nature 346 (6287), 818–822.

Famulok, M., Mayer, G., Blind, M., 2000. Nucleic acid aptamers—from selection *in vitro* to applications *in vivo*. Acc. Chem. Res. 33 (9), 591–599.

Fields, D.S., He, Y., Al-Uzri, A.Y., Stormo, G.D., 1997. Quantitative specificity of the Mnt repressor. J. Mol. Biol. 271 (2), 178–194.

Frech, K., Quandt, K., Werner, T., 1997. Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem. Sci. 22 (3), 103–104.

Ferrieres, L., Francez-Charlot, A., Gouzy, J., Rouille, S., Kahn, D., 2004. FixJ-regulated genes evolved through promoter duplication in *Sinorhizobium meliloti*. Microbiology 150 (7), 2335–2345.

Gerland, U., Moroz, J.D., Hwa, T., 2002. Physical constraints and functional characteristics of transcription factor–DNA interaction. Proc. Natl. Acad. Sci. U.S.A. 99 (19), 12015–12020.

Gilbert, S.D., Batey, R.T., 2005. Riboswitches: natural SELEXion. Cell. Mol. Life Sci. 62 (21), 2401–2404.

Gold, L., 1995. Oligonucleotides as research, diagnostic, and therapeutic agents. J. Biol. Chem. 270 (23), 13581–13584.

Gold, L., Brown, D., He, Y., Shtatland, T., Singer, B.S., Wu, Y., 1997. From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. Proc. Natl. Acad. Sci. U.S.A. 94 (1), 59–64.

Goringer, H.U., Homann, M., Zacharias, M., Adler, A., 2006. RNA aptamers as potential pharmaceuticals against infections with African trypanosomes. Handbook Exp. Pharmacol. 2006 (173), 375–393.

He, Y.Y., Stockley, P.G., Gold, L., 1996. *In vitro* evolution of the DNA binding sites of *Escherichia coli* methionine repressor. MetJ. J. Mol. Biol. 255 (1), 55–66.

Hermann, T., Patel, D.J., 2000. Adaptive recognition by nucleic acid aptamers. Science 287 (5454), 820–825.

Huttenhofer, A., Vogel, J., 2006. Experimental approaches to identify noncoding RNAs. Nucleic Acids Res. 34 (2), 635–646.

Irvine, D., Tuerk, C., Gold, L., 1991. SELEXION. Systematic evolution of ligands by exponential enrichment with integrated optimization by nonlinear analysis. J. Mol. Biol. 222 (3), 739–761.

Jagannathan, V., Roulet, E., Delorenzi, M., Bucher, P., 2006. HTPSELEX—a database of high-throughput SELEX libraries for transcription factor binding sites. Nucleic Acids Res. 34, D90–D94.

Jayasena, S.D., 1999. Aptamers: an emerging class of molecules that rival antibodies in diagnostics. Clin. Chem. 45 (9), 1628–1650.

Jones, S., Daley, D.T., Luscombe, N.M., Berman, H.M., Thornton, J.M., 2001. Protein–RNA interactions: a structural analysis. Nucleic Acids Res. 29 (4), 943–954.

Kainz, S., Czaja, R., Greiner-Stoffele, T., Hahn, U., 2006. Selection of RNase-resistant RNAs. Handbook Exp. Pharmacol. 2006 (173), 447–455.

Kim, S., Shi, H., Lee, D.K., Lis, J.T., 2003. Specific SR protein-dependent splicing substrates identified through genomic SELEX. Nucleic Acids Res. 31 (7), 1955–1961.

Kopylov, A.M., Spiridonova, V.A., 2000. Combinatorial chemistry of nucleic acids: SELEX. Mol. Biol. (Mosk.) 34 (6), 1097–1113.

Kusser, W., 2000. Chemically modified nucleic acid aptamers for *in vitro* selections: evolving evolution. J. Biotechnol. 74 (1), 27–38.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C., 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262 (5131), 208–214.

Liu, J., Stormo, G.D., 2005. Combining SELEX with quantitative assays to rapidly obtain accurate models of protein–DNA interactions. Nucleic Acids Res. 33 (17), e141.

Lorsch, J.R., Szostak, J.W., 1994. *In vitro* evolution of new ribozymes with polynucleotide kinase activity. Nature 371 (6492), 31–36.

Magee, J., Warwicker, J., 2005. Simulation of non-specific protein–mRNA interactions. Nucleic Acids Res. 33 (21), 6694–6699.

Man, T.K., Stormo, G.D., 2001. Non-independence of Mnt repressor operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res. 29 (12), 2471–2478.

Nagai, K., 1996. RNA–protein complexes. Curr. Opin. Struct. Biol. 6 (1), 53–61.

O'Flanagan, R.A., Paillard, G., Lavery, R., Sengupta, A.M., 2005. Non-additivity in protein–DNA binding. Bioinformatics 21 (10), 2254–2263.

Oliphant, A.R., Brandl, C.J., Struhl, K., 1989. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol. Cell. Biol. 9 (7), 2944–2949.

Peng, W., Gerland, U., Hwa, T., Levine, H., 2003. Dynamics of competitive evolution on a smooth landscape. Phys. Rev. Lett. 90 (8), 088103.

Ponomarenko, J.V., Orlova, G.V., Ponomarenko, M.P., Lavryushev, S.V., Frolov, A.S., Zybova, S.V., Kolchanov, N.A., 2000. SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. Nucleic Acids Res. 28 (1), 205–208.

Ravelet, C., Grosset, C., Peyrin, E., 2006. Liquid chromatography, electrochromatography and capillary electrophoresis applications of DNA and RNA aptamers. J. Chromatogr. A 1117 (1), 1–10.

Roberts, D.N., Stewart, A.J., Huff, J.T., Cairns, B.R., 2003. The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. Proc. Natl. Acad. Sci. U.S.A. 100 (25), 14695–14700.

Robison, K., McGuire, A.M., Church, G.M., 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol. 284 (2), 241–254.

Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., Bucher, P., 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat. Biotechnol. 20 (8), 831–835.

Salgado, H., et al., 2004. RegulonDB (Version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. Nucleic Acids Res. 32, D303–D306.

Sarai, A., Takeda, Y., 1989. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proc. Natl. Acad. Sci. U.S.A. 86 (17), 6513–6517.

Schneider, D., Gold, L., Platt, T., 1993. Selective enrichment of RNA species for tight binding to *Escherichia coli* rho factor. FASEB J. 7 (1), 201–207.

Sengupta, A.M., Djordjevic, M., Shraiman, B.I., 2002. Specificity and robustness of transcription control networks. Proc. Natl. Acad. Sci. U.S.A. 99 (4), 2072–2077.

Shi, H., Hoffman, B.E., Lis, J.T., 1997. A specific RNA hairpin loop structure binds the RNA recognition motifs of the *Drosophila* SR protein B52. Mol. Cell. Biol. 17 (5), 2649–2657.

Shimada, T., Fujita, N., Maeda, M., Ishihama, A., 2005. Systematic search for the Cra-binding promoters using genomic SELEX system. Genes Cells 10 (9), 907–918.

Shtatland, T., Gill, S.C., Javornik, B.E., Johansson, H.E., Singer, B.S., Uhlenbeck, O.C., Zichi, D.A., Gold, L., 2000. Interactions of *Escherichia coli* RNA with bacteriophage MS2 coat protein: genomic SELEX. Nucleic Acids Res. 28 (21), e93.

Singer, B.S., Shtatland, T., Brown, D., Gold, L., 1997. Libraries for genomic SELEX. Nucleic Acids Res. 25 (4), 781–786.

Stormo, G.D., Fields, D.S., 1998. Specificity, free energy and information content in protein–DNA interactions. Trends Biochem. Sci. 23 (3), 109–113.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. Bioinformatics 16 (1), 16–23.

Takeda, Y., Sarai, A., Rivera, V.M., 1989. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc. Natl. Acad. Sci. U.S.A. 86 (2), 439–443.

Tombelli, S., Minunni, M., Mascini, M., 2005. Analytical applications of aptamers. Biosens. Bioelectron. 20 (12), 2424–2434.

Tuerk, C., Gold, L., 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249 (4968), 505–510.

Ulrich, H., 2005. DNA and RNA aptamers as modulators of protein function. Med. Chem. 1 (2), 199–208.

Ulrich, H., Trujillo, C.A., Nery, A.A., Alves, J.M., Majumder, P., Resende, R.R., Martins, A.H., 2006. DNA and RNA aptamers: from tools for basic research towards therapeutic applications. Comb. Chem. High. Throughput. Screen. 9 (8), 619–632.

Vant-Hull, B., Payano-Baez, A., Davis, R.H., Gold, L., 1998. The mathematics of SELEX against complex targets. J. Mol. Biol. 278 (3), 579–597.

Vater, A., Jarosch, F., Buchner, K., Klussmann, S., 2003. Short bioactive Spiegelmers to migraine-associated calcitonin gene-related peptide rapidly identified by a novel approach: tailored-SELEX. Nucleic Acids Res. 31 (21), e130.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270 (5235), 484–487.

Wei, G.H., Liu, D.P., Liang, C.C., 2004. Charting gene regulatory networks: strategies, challenges and perspectives. Biochem. J. 381 (1), 1–12.

Wen, J.D., Gray, D.M., 2004. Selection of genomic sequences that bind tightly to Ff gene 5 protein: primer-free genomic SELEX. Nucleic Acids Res. 32 (22), e182.

White, R.R., Sullenger, B.A., Rusconi, C.P., 2000. Developing aptamers into therapeutics. J. Clin. Invest. 106 (8), 929–934.

Wilson, D.S., Szostak, J.W., 1999. *In vitro* selection of functional nucleic acids. Annu. Rev. Biochem. 68, 611–647.

Wingender, E., et al., 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 29, 281–283.

Winter, R.B., Berg, O.G., von Hippel, P.H., 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli* lac repressor–operator interaction: kinetic measurements and conclusions. Biochemistry 20 (24), 6961–6977.

Wright, W.E., Binder, M., Funk, W., 1991. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Mol. Cell. Biol. 11 (8), 4104–4110.

Yagura, M., Itoh, T., 2006. The Rep protein binding elements of the plasmid ColE 2-P9 replication origin. Biochem. Biophys. Res. Commun. 345 (2), 872–877.

Yeung, K.Y., Medvedovic, M., Bumgarner, R.E., 2004. From co-expression to co-regulation: how many microarray experiments do we need? Genome Biol. 5 (7), R48.