Published in final edited form as:
Anal Chem. 2013 November 19; 85(22): 10761–10770. doi:10.1021/ac401875h.

Tracking the Emergence of High Affinity Aptamers for rhVEGF$_{165}$ During CE-SELEX Using High Throughput Sequencing

Meng Jing and **Michael T. Bowser***
Department of Chemistry, University of Minnesota, 207 Pleasant Street SE, Minneapolis, Minnesota, 55455, United States

Abstract

CE-SELEX is a powerful technique for isolating aptamers for various targets, from large proteins to small peptides with molecular weights of several kilodaltons. One of the unique characteristics of CE-SELEX is the relatively high heterogeneity of the ssDNA pools that remains even after multiple rounds of selection. Enriched sequences or highly abundant oligonucleotide motifs are rarely reported in CE-SELEX studies. In this work, we employed 454 pyrosequencing to profile the evolution of an oligonucleotide pool through multiple rounds of CE-SELEX selection against the target rhVEGF$_{165}$. High throughput sequencing allowed up to 3×10^4 sequences to be obtained from each selected pool and compared to the unselected library. Remarkably, the highest abundance contiguous sequence (contig) was only present in 0.8% of sequences even after four rounds of selection. Closer analyses of the most abundant contigs, the top 1000 oligonucleotide fragments and even the eight original FASTA files showed no evidence of prevailing motifs in the selected pools. The sequencing results also provided insight into why many CE-SELEX selections obtain pools with reduced affinities after many rounds of selection (typically >4). Preferential amplification of a particular short PCR product allowed this non-binding sequence to overtake the pool in later rounds of selection suggesting that further refinement of primer design or amplification optimization is necessary. High affinity aptamers with 10^{-8} M dissociation constants for rhVEGF$_{165}$ were identified. The affinities of the higher abundance contigs were compared with aptamers randomly chosen from the final selection pool using affinity capillary electrophoresis (ACE) and fluorescence polarization (FP). No statistical difference in affinity between the higher abundance contigs and the randomly chosen aptamers was observed, supporting the premise that CE-SELEX selects a uniquely heterogeneous pool of high affinity aptamers.

INTRODUCTION

Aptamers are ssDNA or RNA molecules that bind specific targets with high affinity.[1] They have advantages over antibodies in terms of small size, low immunogenicity and facile synthesis, labeling, and storage. Aptamers have been employed broadly in both basic research,[2] and clinical diagnoses and therapies.[3] Aptamers are generated via an *in vitro* process called Systematic Evolution of Ligand by EXponential enrichment (SELEX).[1] A combinatorial starting library containing as many as 10^{15} random ssDNA or RNA sequences is first incubated with the target of interest, and then subjected to partitioning. The binding fraction is collected for PCR amplification and generation of a new pool for the next round of selection. This cycle is repeated until the affinity of the selected oligonucleotide pool

*Corresponding Author bowser@umn.edu.

Supporting Information Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

plateaus. Compared with conventional selection methods such as affinity chromatography and nitrocellulose membrane filtration, capillary electrophoresis (CE) offers a number of advantages.[4] It simplifies the selection by eliminating the need for target immobilization and counter selections. The selection efficiency, which can be defined as the ratio of complexes' collection percentage over unbound sequences' collection percentage, can be improved by two orders of magnitude.[5] As a result, this highly efficient, free solution based separation technique greatly shortens the time needed to obtain high affinity aptamers from 8 ~ 15[6,7] rounds to 2 ~ 4 rounds,[4,7] or even a single round.[5]

One distinctive characteristic of CE-SELEX that contrasts with other SELEX approaches is the heterogeneity retained in the selected pools. Although enriched sequences and conserved oligonucleotides (motifs) were identified in one report after 9 rounds of CE-SELEX selection;[8] to the best of our knowledge, other studies have not identified high abundance sequences or motifs.[4,7,9,10] This is balanced by the observation that CE-SELEX generates pools with a high fraction of sequences with high affinity for the target, suggesting that a higher number of high affinity aptamers are isolated. Clearly, the number of sequences generated via cloning (typically 30-100) is not sufficient to fully characterize the heterogeneity of these selected pools.

Another interesting yet puzzling characteristic of CE-SELEX is that in a number of studies the affinities of the pools decreased beyond a certain selection cycle.[4,9] Tracking the evolution of the pools through the CE-SELEX process with improved sequencing data may provide insight into the origin of this phenomena. Although computational simulations on evolution of aptamers have been proposed,[11] few studies have been reported on tracking the real evolution profiles of individual aptamers. Therefore, a comprehensive study of aptamer emergence with experimental sequencing data will allow us to look into CE-SELEX process with more detail and may shed light on potential improvements and applications.

Revolutionary Next Generation Sequencing (NGS) technologies[12] quickly produce enormous amounts of sequence data at low unit prices when compared to "first generation" Sanger sequencing. These new technologies have given impetus to explosively growing research in *de novo* assemblies of genomes,[13] transcriptome sequencing,[14] and gene discovery in metagenomics.[15] More interestingly to us, NGS has also been used in several SELEX studies, playing powerful roles in Genomic-SELEX[16] and Cell-SELEX[17] to identify aptamers, and to help generate higher affinity aptamers in earlier selection rounds by monitoring the rate of enrichment of individual sequences in microfluidic-SELEX.[18] 454 pyrosequencing is an example of a NGS technology that allows massively parallel sequencing of as many as millions of sequences.[19,20] Briefly, a single stranded template DNA library is first extended to incorporate the adaptor primer and then immobilized onto beads for emulsion PCR (emPCR). Millions of identical copies of the same sequence are amplified onto individual beads. These beads are layered onto a PicoTiterPlate (PTP) device designed to load one bead per well. Sequencing-by-synthesis (pyrosequencing) for each bead is performed in parallel, generating as many as several million reads per run.

Vascular endothelial growth factor (VEGF) is a signaling protein involved in angiogenesis. Over expression of VEGF is associated with many diseases such as age related macular degeneration, rheumatoid arthritis and solid cancers.[21] The first aptamer based drug Macugen was isolated as a $VEGF_{165}$ inhibitor. RNA aptamers[22] and directly selected/ post-selection modified 2′-O-methyl RNA aptamers[23] for $VEGF_{165}$ have been developed using conventional SELEX. DNA aptamers have also been selected against $VEGF_{165}$[24,25] using conventional SELEX. CE-SELEX has been used to select aptamers for a 32 aa peptide fragment of VEGF[26]. From these studies, selected pools converged on several sequences and motifs when conventional SELEX methods were applied, but high diversity remained

when CE was chosen as the partitioning method. With this knowledge in mind, we decided to use $VEGF_{165}$ as the target and 454 pyrosequencing as the characterization method in our study to probe the evolution process of aptamers selected by CE-SELEX.

EXPERIMENTAL

Chemicals

Recombinant Human Vascular Endothelial Growth Factor 165 ($rhVEGF_{165}$, MW 44 kDa) was purchased from ProSpec Protein Specialists, Inc. (East Brunswick, NJ). Nuclease free H_2O, forward primer 5′-FAM-AGC AGC ACA GAG GTC AGA TG-3′, reverse primer 5′-Biotin-TTC ACG GTA GCA CGC ATA GG, fusion primer A 5′-GCC TCC CTC GCG CCA TCA GAG CAG CAC AGA GGT CAG ATG-3′, fusion primer B 5′-GCC TTG CCA GCC CGC TCA GTT CAC GGT AGC ACG CAT AGG-3′, random ssDNA library 5′-FAM-AGC AGC ACA GAG GTC AGA TG $(N)_{40}$ CCT ATG CGT GCT ACC GTG AA-3′, and individual aptamers were purchased from Integrated DNA Technologies, Inc. (Coralville, IA). 100 mM dNTP's and 25 bp DNA ladder were purchased from Invitrogen, Inc. (Carlsbad, CA). *Taq* DNA polymerase and ThermoPol buffer were purchased from New England BioLabs, Inc. (Ipswich, MA). Blue/orange gel loading dye was purchased from Promega Corp. (Madison, WI). Streptavidin agarose resin was purchased from Thermo Scientific, Inc. (Waltham, MA). Acetic acid (CH_3COOH, 99.7%, Mallinckrodt Baker, Saint Louis, MI), boric acid for electrophoresis (H_3BO_3, Sigma-Aldrich, Saint Louis, MO), DL-Dithiothreitol (DTT, MP Biomedicals, Irvine, CA), ethylenediaminetetraacetic acid disodium salt dihydrate (EDTA, 99.0%-101.0%, Sigma-Aldrich, Saint Louis, MO), glycine ($C_2H_5NO_2$, 99%, Sigma-Aldrich, Saint Louis, MO), 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES, 99%, Alfa Aesar, Ward Hill, MA), magnesium chloride 6-hydrate ($MgCl_2{\cdot}6H_2O$, 99.8%, Mallinckrodt Baker, Saint Louis, MI), potassium phosphate, monobasic, (KH_2PO_4, 99.9%, Mallinckrodt Baker, Saint Louis, MI), sodium chloride (NaCl, 99.0%, Mallinckrodt Baker, Saint Louis, MI), sodium hydroxide (NaOH, 97%, Sigma-Aldrich, Saint Louis, MO), and tris(hydroxymethyl) aminomethane (tris, 99.8%, Sigma-Aldrich, Saint Louis, MO) were used without further purification. Binding buffer, CE separation buffer, and streptavidin buffer were prepared in nuclease free H_2O, and 0.5×TBE buffer was prepared in Milli-Q H_2O from a Milli-Q water purification system (Millipore Corp., Bedford, MA). Binding buffer consisted of 20 mM HEPES, 10 mM NaCl, and 0.25 mM DTT at pH 7.2 (pH was adjusted using a 1M NaOH solution). CE separation buffer consisted of 25 mM tris, 192 mM glycine, and 5 mM KH_2PO_4 (TGK) at pH 8.3. Streptavidin binding buffer consisted of 10 mM tris, 50 mM NaCl, and 1mM EDTA at pH 7.5. 0.5 × TBE buffer consisted of 45 mM tris, 45 mM boric acid, and 1 mM EDTA at pH 8.3. All buffers were filtered through a 0.2 μm posidyne® membrane filter before use.

Capillary Electrophoresis Selections

All CE selections were performed on a P/ACE™ MDQ Capillary Electrophoresis System from Beckman Coulter, Inc. (Fullerton, CA). An argon ion laser (Beckman coulter, Inc., Fullerton, CA) with a 488 nm line was used for excitation and fluorescence emission was collected at 520 nm. A 50 cm long (40 cm to the detector) uncoated fused silica capillary with an inner diameter of 50 μm and outer diameter of 360 μm (Polymicro Technologies, Phoenix, AZ) was prepared each day before separation and collection. The capillary was rinsed with 0.15 M NaOH and then CE separation buffer at 30 psi for 5 min before the first separation of the day, and 1 min for the following separations. ssDNA library dissolved in binding buffer was heated to 72 °C for 5 min, allowed to cool to room temperature, and then incubated with VEGF for 30 min. In the first round of selection, the mixture contained 10 μL of the random ssDNA library at a concentration of 100 μM and VEGF at 10 nM. The mixture was hydrodynamically injected at 1 psi for 4 seconds. Based on Hagen-Poiseuille

law and assuming the dynamic viscosity of the binding buffer is approximately the same as that of water at 25 °C, the injection volume can be estimated as 9 nL:

$$V=Qt=\frac{\Delta P\pi d^4 t}{128\mu L} \quad (1)$$

in which V is the injection volume, t is the injection time, ΔP is the pressure drop along the capillary, μ is the dynamic viscosity of the injected buffer, d is the inner diameter of the capillary, and L is the total length of the capillary. This means for each injection approximately 5×10^{11} ssDNA molecules and 5×10^{7} VEGF molecules were loaded onto the capillary. The separation was performed at 30 kV under normal polarity at 25 °C. The cutoff time for fraction collection was calculated according to the ratio of the effective capillary length to the total capillary length. Complexes migrating 45 sec before the rise of the unbound ssDNA peak were collected at the capillary outlet into a vial containing 48 μL separation buffer for PCR amplification. In the following rounds of selection, the highest concentration ssDNA available was used with an aliquot retained for binding experiments (~1 μM), and VEGF was maintained at 10 nM.

PCR Amplification

A total volume of 943 μl PCR master mixture was prepared with the recipe of 1 mM dNTP's, 7.5 mM $MgCl_2$, 500 nM forward and reverse primers, and 1× ThermoPol buffer. The mixture was aliquoted into nine PCR vials. 6 μL of binding buffer was added to the first vial as a negative control and 6 μL of the collected CE fraction was added to each of the remaining eight vials. All the PCR vials were loaded onto a Mastercycler® (Eppendorf, Enfield, CT), heated to 94 °C for 1 min, and paused at 94 °C. 5 units *taq* polymerase was added into each of the 9 vials. 23 cycles of PCR were performed at 94 °C for 30 s for denaturation, 55 °C for 30 s for annealing, and 72 °C for 20 s for extension. A final extension at 72 °C for 5 min was carried out after the 23rd cycle. An agarose gel (2%) electrophoresis of the PCR product was performed to observe the desired product at 80 bp and possible contamination in the negative control or byproducts at other lengths. A purification step using streptavidin agarose resin and a chromatography column was used to collect the single strands with FAM labels, after which standard ethanol precipitation was used to isolate the final product as previously reported[4,9]. The dry product was then dissolved in 10 μL binding buffer for the next round of selection and affinity characterization.

Sequencing and Bioinformatics

As a control 1 fmol of the original random ssDNA library was PCR amplified and purified as described above. Aliquots from the selected CE-SELEX pools and a 1,000-fold dilution of the control sequences were subjected to an additional PCR step to incorporate the fusion primer. PCR was performed under the same conditions as described above, except that fusion primer A and B were used to add the adaptors necessary for 454 sequencing. PCR products were then loaded onto 2% agarose gels and electrophoreses was carried out at 100 V for 3.5 hours. Gel images were taken to ensure there was no contamination in all eight reactions. The 80 bp bands were carefully cut out of the gel while being monitored on a DNA safe Visi-Blue™ transilluminator (UVP, Inc., Upland, CA). These 80 bp products were extracted using QIAquick® gel extraction kit (Qiagen, Inc., Gaithersburg, MD) and stored in TE buffer of pH 7.0 at −20 °C. 454 sequencing was performed at the BioMedical Genomics Center at the University of Minnesota. Sequencing results were exported as eight FASTA files, one for each selected pool and another for the control sequences.

Five of the eight FASTA files contained more than 20,000 sequences, making multiple sequence alignment or motif identification very time consuming. Thus, an in house programmed code in Linux was used to quickly identify enriched sequences/motifs. Briefly, the desired products with lengths from 78 to 82 were extracted for analysis. The primer regions were removed before analysis. An oligonucleotide fragment with a length of n nt (n ranged from 5 to 15) was then extracted from the 1st register position to the nth position. Then the second oligonucleotide was extracted from the 2nd to the (n + 1)th register position, and the process continued to the last oligonucleotide from the (41 − n)th to the 40th register position. The same analysis moved on to the second sequence in the file and progressed through each sequence. While extracting, every oligonucleotide fragment was recorded, and the count number was added if a same one was found later in the analysis. When the program finished, all the oligonucleotides were exported in order of decreasing abundance. Enriched sequences were identified by piecing together abundant and overlapping fragments into complete 40 nt sequences. The results were confirmed by CodonCode Aligner3.7.1 (CodonCode Corp., Bedham, MA), which identified the top three enriched sequences when analyzing each file divided into subgroups of ~3000 sequences. Online software MEME4.7.0[27,28] and ClustalW2.0[29] were also used in order to find motifs both in the short byproducts and the desired products.

K_d Measurements by Affinity Capillary Electrophoresis (ACE) and Fluorescence Polarization (FP)

ACE experiments were performed using the same MDQ CE system under the same separation conditions. Selected pools/aptamers with a fixed concentration in the range of 1 nM to 10 nM were heated to 72 °C for 5 min, and then allowed to cool to room temperature. 10 μL of these samples were then incubated with 10 μL of 0 nM to 750 nM VEGF. The mixtures were then injected onto the capillary from the lowest to the highest VEGF concentration. Electropherograms were recorded and analyzed using 32 Karat software (Beckman Coulter, Inc., Fullerton, CA). The peak heights of the free aptamer were used to calculate the bound fractions and the K_d value was obtained using a nonlinear fitting of equation 2[30] in Origin 8.1 (OriginLab, Inc., Northampton, MA):

$$f_a = \frac{C}{1+\frac{K_d}{\left([VEGF]_t - 0.5\left([A]_t + [VEGF]_t + K_d - \left(([A]_t + [VEGF]_t + K_d)^2 - 4[A]_t[VEGF]_t\right)^{0.5}\right)\right)}} \quad (2)$$

in which f_a, $[VEGF]_t$, $[A]_t$, and C are the bound fraction of aptamer, total VEGF concentration, total aptamer concentration, and the maximum bound fraction, respectively. $([VEGF]_t - 0.5([A]_t + [VEGF]_t + K_d - (([A]_t + [VEGF]_t + K_d)^2 - 4[A]_t[VEGF]_t)^{0.5}))$ is the concentration of unbound VEGF.

Fluorescence polarization (FP) experiments were performed on a Synergy™ 2 microplate reader (BioTek Instruments, Inc., Winooski, VT). Samples were prepared in the same manner as in the ACE experiments. The 20 μL mixtures were loaded into a corning 3540 microplate (Corning Inc., Corning, NY). Gen 5™ software (BioTek^R Instruments, Inc., Winooski, VM) was used to record parallel and perpendicular intensities (λ_{ex} = 485 ± 20 nm, λ_{em} = 528 ± 20 nm) and calculate polarization:

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \quad (3)$$

Each sample was measured three times and all data were used in the K_d analysis. The same equation and software were used to determine K_d as in ACE, except the bound fraction which was determined according to:

$$f_a = \frac{P - P_o}{P_m - P_o} \quad (4)$$

where P, P_o, and P_m are the polarization of one sample, the free aptamer, and the bound aptamer, respectively. The overall fluorescence intensities of the samples were also monitored and the bound fractions were corrected using equations 5 and 6 as needed[31] :

$$Q_m = \frac{I_m}{I_o} I_o \quad (5)$$

$$f_a' = \frac{f_a}{1 + Q_m (1 - f_a)} \quad (6)$$

where I_m, I_o, Q_m and f_a' are the fluorescence intensities of bound aptamer, free aptamer, the fluorescence intensity enhancement factor, and the corrected bound fraction, respectively.

RESULTS AND DISCUSSION

Affinity of the Selected Pools during CE-SELEX

Prior to performing selections, the affinity of the initial ssDNA library for rhVEGF165 was evaluated using two orthogonal methods: affinity capillary electrophoresis (ACE) and fluorescence polarization (FP). Limited by the concentration of our VEGF stock solutions, we were only able to assess affinity up to 1 μM. The random ssDNA library had no apparent affinity toward VEGF even at the highest VEGF concentration, suggesting a K_d value significantly higher than 1 μM. This is not surprising considering DNA aptamers for VEGF have been selected with a few hundred nanomolar K_d values.[24] Seven cycles of CE selection were performed. The incubated mixture in cycle one consisted of 100 μM library and 10 nM VEGF resulting in ~5×10^{11} ssDNA sequences and ~5×10^7 VEGF molecules injected per separation. Three CE separations were performed for each round of selection to increase the number of binding sequences collected. Most of the purified ssDNA product was reserved for the subsequent round of selection with small aliquot allocated for K_d measurements. As shown in Fig 1, the affinity of the bulk pool improved by over an order of magnitude after a single round of CE-SELEX selection, with the K_d values of the pool estimated to be 50 ± 60 nM and 120 ± 80 nM by ACE and FP, respectively. The affinity of the pool for VEGF continued to improve modestly until rounds 6 and 7, after which fluctuations were observed, a phenomenon reported in previous CE-SELEX studies.[4,9]

Overview of Sequencing Results

After seven rounds of CE-SELEX selection, an aliquot from each selected pool was PCR amplified to incorporate the adaptor primers. The products were gel purified and recovered for 454 sequencing. A control sample was generated from the starting library (as described in the experimental section) and sequenced for comparison with the selected pools. Figure 2 summarizes the composition of the sequencing results over the seven selection rounds. The same volume and procedure was used for each pool allowing variations in quantity and quality of the samples after each round of CE-SELEX selection to be directly compared. After the first round of selection, both the number of sequences and the average sequence length were very similar to those observed for the control. Both the number and the length of the sequences decreased mildly with increased rounds of selection up to round 5. Further

selection in rounds 6 and 7 resulted in a sharp decrease in both number and quality of sequences.

Analysis of the Short Sequence Byproducts

The distribution of sequence lengths for each selected pool was plotted to further characterize the progression of the pool through multiple rounds of selection (see Figure 3). The frequency distributions were normalized to the abundance of the 80 nt sequences to account for variations in the number of sequences obtained in each round (see Figure 2). The most abundant length for all the pools was 80 nt, which was expected since the starting library had a length of 80 nt. Very few sequences with lengths longer than 82 nt were observed in any of the selected pools. However, an increased abundance of short sequences was observed with increasing rounds of selection. No significant shorter than expected sequences were observed in the control or the pool after a single round of selection. An obvious peak centered at 47 nt emerged from rounds 2 through 5 with the normalized abundances approaching 10%. In rounds 6 and 7 a broader distribution of short sequences was observed suggesting the emergence of a wider range of undesired byproducts. The length distribution of the short sequences in these rounds ranged from 42 nt to as high as 65 nt, with the normalized abundances of several sequence lengths reaching over 20%.

The short byproducts with highest abundance lengths were further characterized using MEME to assess heterogeneity and identify motifs. The normalized abundances of the short sequences at a specific length, the identified motifs, and the normalized abundances of these motifs are summarized in Table 1. The control sample was treated exactly the same as pool 1 (i.e. PCR amplification, purification, a second PCR to incorporate adaptor primers, and 454 sequencing) with the exception of incubation with VEGF and selection using CE. The undesired short products only made up a small fraction of pools for the control and after one round of selection, less than 0.3% the abundance of the 80 nt sequences. The motif identified in the control showed a low level of consensus (see supporting information, Figure S1. (I)) and was found to be randomly distributed without preference to a particular register position. The normalized abundance of sequences containing this motif was only 0.035% (as determined by MEME), compared to 0.157% for all the 45 nt sequences, suggesting a high degree of heterogeneity remained within this subset of the overall pool. After one round of selection, a 10 nt motif was identified with a very high consensus level (see supporting information, Figure S1. (A)). This motif was found in two thirds of the 50 nt short sequences and 93% of the time was located in the same register position, suggesting that majority of this byproduct was from the same origin. Interestingly, although this motif also showed up in the short byproducts in subsequent rounds of selection, it was not the dominant short sequence byproduct. From rounds 2 through 6, the most abundant short byproducts were 47 nt long, shared the same motif, made up the majority of the short sequence byproducts and for the most part occurred at the same register position. Again, this suggests that the majority of the short sequence byproducts shared a common source. By round 7 the distribution of short byproducts became more diverse with three observed length populations (54 nt, 46 nt, and 42 nt) with normalized abundances over 20%. The motif identified in the 46 nt sequences was similar to the motif found in rounds 2 through 6. 42 nt sequences were primer dimers with a random 2 nt region in the middle, and 54 nt sequences shared a newly discovered motif. These motifs identified in the pools from rounds 1 through 7 were also observed in some short byproducts at other lengths and the few long byproducts.

Although it is difficult to determine an explicit mechanism for the formation of these short byproducts, these motifs share a common characteristic of containing more GC than AT, suggesting failure to amplify GC-rich sequences as a contributing factor. It is also worth pointing out that careful gel-electrophoresis separation and extraction were performed to eliminate undesired products before 454 sequencing. A gel image taken after purification

clearly showed single bands at 80 bp (see supporting information, Figure S2). However, the short byproducts reached as high as 28% of the abundance of sequences with the proper length. After the adaptor primers were incorporated our sequence lengths were well within the tolerances of the 454 instrument, suggesting that the sequencing error rate should be below 1%.[19] Thus we can conjecture that the short sequences have much higher amplification efficiency and were preferably amplified during emPCR. The amplification preference of the short sequences directly impacted the progression of the CE-SELEX selection beyond round 5 and the bulk affinities of the pools were not further improved or even deteriorated. Further studies are needed to determine if redesign of primer sequences can limit the emergence of these short byproducts or if it is an unavoidable consequence of performing multiple rounds of PCR on structured aptamer sequences.

Analysis of Sequences with the Proper Length

454 sequencing generated up to ~3×10^4 sequences for each selected pool, among which the most abundant sequences were the desired products with lengths within the range of 78 nt to 82 nt (see Figures 2 and 3). This several orders more sequence data than typically produced by cloning, which generally produces 30-100 sequences, greatly expanding our ability to characterize the heterogeneity of the CE-SELEX pools. However, the massive amount of data generated brings additional challenges. Commonly available sequence-alignment or motif-finding programs are not designed to accommodate this much sequence data and quickly become prohibitively time consuming. For example, the run time of MEME is proportional to the size of the data set squared, requiring months to analyze a 30,000 sequence file.[27] As a result, we decided to employ several complementary, less computationally intensive approaches to identify common sequence fragments and motifs in the selected pools.

In house software was programmed in Linux to rank the abundance of oligonucleotide fragments present within the 78 and 82 nt sequences amongst all of the selected pools (see experimental section, sequencing and bioinformatics for details). The 1000 most abundant oligonucleotide fragments with lengths ranging from 5 nt through 15 nt were exported in descending order of abundance (see the top 100 in supporting information, Table S1). The probability of an n nt long sequence fragment randomly occurring in a ssDNA pool with a 40 nt random region equals $0.25^n \times (41 - n) \times 100\%$. The logarithms of these theoretical values are displayed in Figure 4 (a). A particular oligonucleotide fragment would be expected to be present at these theoretical abundances in a random sequence pool. Fragments present at abundances higher than the theoretical values suggest enrichment has occurred through the CE-SELEX process. Figure 4b illustrates how the abundance of the most frequently observed oligonucleotide fragments of different lengths progressed through the rounds of CE-SELEX selection. To clearly emphasize the magnitude of enrichment, abundances were normalized to the theoretically predicted abundances in a random oligonucleotide pool and then displayed as a logarithm to facilitate presentation of the large range of values. The most abundant 5 nt motif TTTTG had an abundance similar to that expected in a randomly generated sequence, giving rise to a value near the non-enriched value of 0. This suggests that there were no 5 nt oligonucleotides present in the pools that were preferably enriched. The most abundant 6 nt and 7 nt oligonucleotides, however, had the highest logarithm values of 0.52 and 0.91 at cycle 5, respectively, corresponding to 3.3:1 and 8.1:1 ratios, indicating modest enrichment. Motifs with lengths ranging from 8 nt through 15 nt, were significantly enriched after a single round of selection with maximum enrichment occurring in round 4.

For comparison, the abundances of the top oligonucleotide fragments found in the control sample matched well with the theoretical random values for motifs ranging from 5 nt to 9 nt (see Figure 4 (a)), confirming that there was little bias in the starting library. The difference

between the theoretical abundances and those observed experimentally in the control only became significant for motifs longer than 9 nt. This is because the probability of a motif longer than 9 nt being found in a 30,000 sequence pool is less than one. However, practically the most abundant motif (larger than 9 nt) has to occur at least once.

As shown in Figure 4 (b), all the identified motifs shared common fragments (excluding the "TTTTG" motif that demonstrated little to no enrichment). In addition, most of these motifs appeared at the same register position in the desired products. For example, 65% of "GGAGCC" were found as the last six nucleotides in the 40 nt random region in pool 1, while the remaining 35% were randomly distributed in the 40 nt random region with an abundance of 0.17%, compared to the abundances of 0.85% in theory and 0.57% in the control (see Figure 4(a)). Another interesting phenomenon is that this fragment was also found in the short byproducts from pool 2 through pool 6, supporting our premise that the short byproducts were formed as a result of the failed amplification of a GC-rich sequence.

The 1000 most abundant fragments ranging from 5 nt through 15 nt were further analyzed to identify common sequences or contigs. Similar to approaches used in genomic sequencing, contigs were identified by aligning motifs to construct common 40 nt sequences. As expected, no meaningful results were obtained when aligning the 5 nt motifs since the abundance of these fragments were no larger than those expected for a random pool. All of the remaining files converged on similar results and contigs were discovered. We also obtained the top 1000 abundant oligonucleotide fragments at each length from 5 nt through 15 nt observed in the control file, and performed the same analysis. The most abundant contig in the control showed up seven times, corresponding to an abundance of 2.33×10^{-2} % in the desired products. This slight enrichment was likely the result of slight bias through the multiple PCR reactions and the purification process. This control threshold was useful for identifying contigs with significant enrichment due to CE-SELEX selections in the other seven pools. Nine contigs were found to have abundances greater than 2.33×10^{-2} % in at least one pool. These contigs are summarized in Table 2 in order of descending abundance as C1 through C9. Not surprisingly, the most abundant motifs (6 nt to 15 nt) shown in Figure 4 are present as fragments of C1. Within these nine contigs, only the top three had abundances that were significantly higher than 2.33×10^{-2} % (S/N = 3). The evolution profiles of the three most abundant contigs are shown in Figure 5. None of the three contigs were identified in the control. C1 had a significant abundance after a single round of selection, with abundance increasing until pool 4, after which a slight decrease was observed in pool 5 and a sharper decrease in pools 6 and 7. This trend is opposite to that of the most abundant short byproducts as shown in Figure 2, confirming the deleterious effect of the emerging short sequence byproducts on the selection process. The remaining two contigs demonstrated similar evolution profiles. It should be noted that the abundance of C1 peaked at less than 1% after 4 rounds of selection, confirming that a relatively high sequence diversity was retained throughout the CE-SELEX selection process. At this abundance it is not surprising that previous CE-SELEX selections have failed to identify common sequences from 30-100 clones.

The top 1000 abundant fragments at 6 nt through 15 nt were further analyzed to search for possible motifs in the selected pools. 78-82 nt sequences were searched for the presence of individual fragments. However, identified fragments either appeared at the same register position or were randomly distributed at abundances approaching the theoretical abundances of a random oligonucleotide. Since the in house programmed code was unable to find mismatched or discontinuous motifs, we divided the original eight pools into subgroups containing ≤ 5,000 sequences and analyzed these subgroups using MEME. This size was chosen as a compromise between run time of the program and representative sampling of the diversity in the original files. Identified motifs with high and medium consensus levels

matched with the previously identified high abundance fragments. CodonCode Aligner3.7.1 was also employed to perform sequence alignment on 3,000-sequence subgroups. Again, no new motifs or contigs were discovered beyond those previously identified. MEME and ClustalW2 were then used to perform alignment on the discovered contigs. Although several motifs were found with medium consensus level using MEME, the results by ClustalW2 clearly showed heterogeneity between the contigs, with the highest similarity score of 42 out of 100, between C2 and C6. The motif regions were highlighted in different colors in Table 3, and the motifs are given in the supporting information (Figure S3).

Aptamer Affinity

To further characterize the aptamers, the three most abundant contigs and another three sequences chosen at random from pool 4 (see Table 2, R1 ~ R3) were synthesized and their affinity for VEGF assessed by both ACE and FP (see Table 3). Figure 6 (a) shows the electropherograms obtained by titrating 5 nM R3 with increasing concentrations of VEGF. The decrease of the free aptamer peak started immediately at 5.9 nM VEGF. At VEGF concentrations above 94 nM a small peak ahead of the free aptamer peak was observed that increased with increasing ligand concentration, consistent with observation of the aptamer-VEGF complex. Figure 6 (b) is the binding curve obtained from the ACE experiment. The K_d estimated using ACE was 24 nM ± 11nM, similar to the 46 nM ± 11 nM estimated using FP (see Figure 6 (c)). Table 3 summarizes the K_d values of the six sequences estimated using both methods. The overall trend is that smaller K_d values were obtained by ACE, even after modification to bound fractions in FP due to fluorescence quenching (see experimental section). However, most K_d ranges from the ACE experiments were not significantly different than those from the FP experiments. The affinities obtained for the individual aptamers were also similar to those of measured for the oligonucleotide pools throughout the selection and were comparable to or better than previously reported DNA aptamers selected against $VEGF_{165}$ or VEGF peptide.[24,25,32] It should be noted that none of the 27 previously reported aptamer sequences for VEGF were observed any of our selected pools. This is not surprising since different primers or target preparations would be expected to significantly affect the aptamer sequences that are selected.

It is important to note that the K_d's for the randomly chosen aptamers are statistically indistinguishable from those of the most abundant sequences. The randomly chosen sequences were assessed to compare the distribution of K_d's for aptamers sampled from across the pool with the K_d's for the most abundant and presumably strongest binding sequences. The fact that no difference was observed suggests that even the low abundance sequences in the selected pool exhibit high affinity for VEGF. Combined with the observed heterogeneity of the pool this confirms that CE-SELEX is capable of producing thousands of high affinity aptamer sequences in as few as 4 rounds of selection.

CONCLUSIONS

In summary, we successfully generated ssDNA aptamers for $rhVEGF_{165}$ using CE-SELEX for the first time. Affinities of abundant and randomly chosen aptamers for VEGF were characterized using two orthogonal methods, ACE and FP, with estimated low nanomolar dissociation constants. High-throughput 454 pyrosequencing generated up to 3×10^4 sequences from each selected pool as well as a control sample, allowing an in-depth study of the evolution profile of CE-SELEX. By combining in house programmed code with online and commercially available software, abundant oligonucleotide fragments and three enriched contigs were identified. Instead of randomly distributed in the 40 nt random region, the most abundant fragments appeared at the same register position, and overlapped each other to extend to full 40 nt contigs. MEME and CodonCode Aligner analyses on 5000/3000-sequence subgroups confirmed these abundant oligonucleotides and contigs,

without discovery of new motifs. Although enriched sequences were identified, the highest abundance observed was 0.8%, indicating that heterogeneity remained high even in the pool with the highest binding affinity. The evolution profile of the undesired short byproducts was also monitored. The abundance of these short sequence byproducts increased dramatically after rounds 6 and 7, contributing to the decrease in abundance of the three binding contigs in these same pools. A closer look into the most abundant short byproducts revealed that they were mistakenly amplified C1. Thus, this study confirms one of the advantages of CE-SELEX, which is to generate highly heterogeneous binding pools containing a diverse set of aptamers beyond the several binding motifs typically obtained in SELEX selections. In contrast, CE-SELEX appears to be more vulnerable to the emergence of short sequence byproducts, limiting the number of selection rounds. Further experiments are necessary to determine if this limitation is best alleviated through improved primer design or PCR optimization.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge Dr. Zhengjing Tu for his help in developing the in-house code to identify enriched oligonucleotides. This research was funded by the National Institutes of Health (R01 GM063533).

REFERENCES

(1). Ellington AD, Szostak JW. Nature. 1990; 346:818–822. [PubMed: 1697402] Tuerk C, Gold L. Science. 1990; 249:505–510. [PubMed: 2200121]

(2). Tombelli S, Minunni A, Mascini A. Biosensors & Bioelectronics. 2005; 20:2424–2434. [PubMed: 15854817]

(3). Nimjee SM, Rusconi CP, Sullenger BA. Annual Review of Medicine. 2005; 56:555–583.

(4). Mendonsa SD, Bowser MT. Journal of the American Chemical Society. 2004; 126:20–21. [PubMed: 14709039] Mosing RK, Mendonsa SD, Bowser MT. Analytical Chemistry. 2005; 77:6107–6112. [PubMed: 16194066] Mendonsa SD, Bowser MT. Journal of the American Chemical Society. 2005; 127:9382–9383. [PubMed: 15984861]

(5). Berezovski M, Drabovich A, Krylova SM, Musheev M, Okhonin V, Petrov A, Krylov SN. Journal of the American Chemical Society. 2005; 127:3165–3171. [PubMed: 15740156]

(6). Wiegand TW, Williams PB, Dreskin SC, Jouvin MH, Kinet JP, Tasset D. Journal of Immunology. 1996; 157:221–230.Schneider DJ, Feigon J, Hostomsky Z, Gold L. Biochemistry. 1995; 34:9599–9610. [PubMed: 7542922] Proske D, Hofliger M, Soll RM, Beck-Sickinger AG, Famulok M. Journal of Biological Chemistry. 2002; 277:11416–11422. [PubMed: 11756401]

(7). Tang JJ, Xie JW, Shao NS, Yan Y. Electrophoresis. 2006; 27:1303–1311. [PubMed: 16518777]

(8). Mallikaratchy P, Stahelin RV, Cao ZH, Cho WH, Tan WH. Chemical Communications. 2006:3229–3231. [PubMed: 17028752]

(9). Mendonsa SD, Bowser MT. Analytical Chemistry. 2004; 76:5387–5392. [PubMed: 15362896]

(10). Cella LN, Sanchez P, Zhong WW, Myung NV, Chen W, Mulchandani A. Analytical Chemistry. 2010; 82:2042–2047. [PubMed: 20136122] Tok J, Lai J, Leung T, Li SFY. Electrophoresis. 2010; 31:2055–2062. [PubMed: 20564698] Williams BAR, Lin LY, Lindsay SM, Chaput JC. Journal of the American Chemical Society. 2009; 131:6330–6331. [PubMed: 19385619] Lin L, Hom D, Lindsay SM, Chaput JC. Journal of the American Chemical Society. 2007; 129:14568–14569. [PubMed: 17985909]

(11). Irvine D, Tuerk C, Gold L. Journal of Molecular Biology. 1991; 222:739–761. [PubMed: 1721092] Sun FZ, Galas D, Waterman MS. Journal of Molecular Biology. 1996; 258:650–660. [PubMed: 8636999] Djordjevic M, Sengupta AM. Physical Biology. 2006; 3:13–28. [PubMed:

16582458] Levine HA, Nilsen-Hamilton M. Computational Biology and Chemistry. 2007; 31:11–35. [PubMed: 17218151]

(12). Shendure J, Ji HL. Nature Biotechnology. 2008; 26:1135–1145.Metzker ML. Nature Reviews Genetics. 2010; 11:31–46.

(13). Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. Nature. 2008; 452:872–U5. [PubMed: 18421352] Li RQ, Zhu HM, Ruan J, Qian WB, Fang XD, Shi ZB, Li YR, Li ST, Shan G, Kristiansen K, Li SG, Yang HM, Wang J. Genome Research. 2010; 20:265–272. [PubMed: 20019144]

(14). Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Nature Methods. 2008; 5:613–619. [PubMed: 18516046]

(15). Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Clinical Chemistry. 2009; 55:856–866. [PubMed: 19264858]

(16). Zimmermann B, Gesell T, Chen D, Lorenz C, Schroeder R. Plos One. 2010; 5:e9169. [PubMed: 20161784]

(17). Bayrac AT, Sefah K, Parekh P, Bayrac C, Gulbakan B, Oktem HA, Tan WH. ACS Chemical Neuroscience. 2011; 2:175–181. [PubMed: 21892384]

(18). Cho M, Xiao Y, Nie J, Stewart R, Csordas AT, Oh SS, Thomson JA, Soh HT. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:15373–15378. [PubMed: 20705898]

(19). Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D. Genome Biology. 2007:8.

(20). Ronaghi M. Genome Research. 2001; 11:3–11. [PubMed: 11156611]

(21). Kim KJ, Li B, Winer J, Armanini M, Gillett N, Phillips HS, Ferrara N. Nature. 1993; 362:841–844. [PubMed: 7683111] Leung DW, Cachianes G, Kuang WJ, Goeddel DV, Ferrara N. Science. 1989; 246:1306–1309. [PubMed: 2479986] Myoken Y, Kayada Y, Okamoto T, Kan M, Sato GH, Sato JD. Proceedings of the National Academy of Sciences of the United States of America. 1991; 88:5819–5823. [PubMed: 1712106] Rosen LS. Cancer Journal. 2001; 7:S120–S128.

(22). Ruckman J, Green LS, Beeson J, Waugh S, Gillette WL, Henninger DD, Claesson-Welsh L, Janjic N. Journal of Biological Chemistry. 1998; 273:20556–20567. [PubMed: 9685413] Jellinek D, Green LS, Bell C, Janjic N. Biochemistry. 1994; 33:10450–10456. [PubMed: 7520755]

(23). Green LS, Jellinek D, Bell C, Beebe LA, Feistner BD, Gill SC, Jucker FM, Janjic N. Chemistry & Biology. 1996; 3:960–960.Burmeister PE, Lewis SD, Silva RF, Preiss JR, Horwitz LR, Pendergrast PS, McCauley TG, Kurz JC, Epstein DM, Wilson C, Keefe AD. Chemistry & Biology. 2005; 12:25–33. [PubMed: 15664512]

(24). Hasegawa H, Sode K, Ikebukuro K. Biotechnology Letters. 2008; 30:829–834. [PubMed: 18175068] Potty ASR, Kourentzi K, Fang H, Jackson GW, Zhang X, Legge GB, Willson RC. Biopolymers. 2009; 91:145–156. [PubMed: 19025993]

(25). Nonaka Y, Sode K, Ikebukuro K. Molecules. 2010; 15:215–225. [PubMed: 20110884] Taylor JN, Darugar Q, Kourentzi K, Willson RC, Landes CF. Biochemical and Biophysical Research Communications. 2008; 373:213–218. [PubMed: 18555799]

(26). Rose CM, Hayes MJ, Stettler GR, Hickey SF, Axelrod TM, Giustini NP, Suljak SW. Analyst. 2010; 135:2945–2951. [PubMed: 20820497]

(27). Bailey TL, Elkan C. Machine Learning. 1995; 21:51–80.

(28). Bailey TL, Williams N, Misleh C, Li WW. Nucleic Acids Research. 2006; 34:W369–W373. [PubMed: 16845028]

(29). Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Nucleic Acids Research. 2003; 31:3497–3500. [PubMed: 12824352] Thompson JD, Higgins DG, Gibson TJ. Nucleic Acids Research. 1994; 22:4673–4680. [PubMed: 7984417]

(30). Hall, KB.; Kranz, JK. Humana Press; Totowa, New Jersy: 1999. p. 109-110.Jing M, Bowser MT. Analytica Chimica Acta. 2011; 686:9–18. [PubMed: 21237304]

(31). Wei AP, Herron JN. Analytical Chemistry. 1993; 65:3372–3377. [PubMed: 8297026]

(32). Rose C, Hayes M, Stettler G, Hichey Scott, Axelrod T, Giustini N, Suljak S. Analyst. 2010

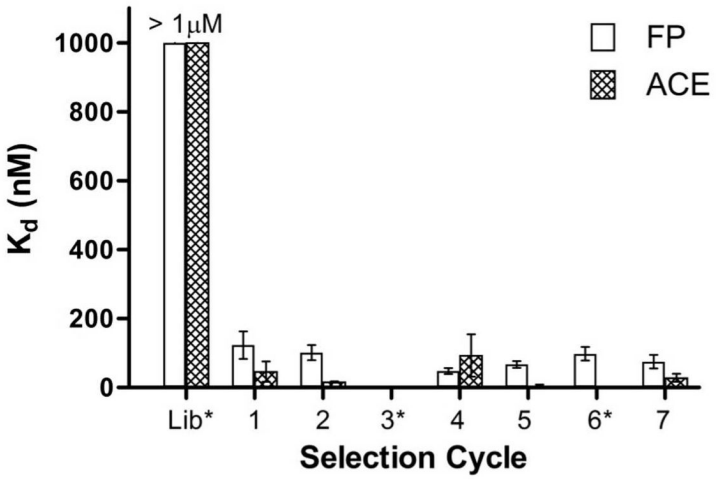

Figure 1.

Binding affinities (K_d values) of the initial random ssDNA library and the pools after CE-SELEX selection for VEGF. K_d values were obtained by both FP and ACE. Error bars represent 95% confidence intervals. Affinity of the initial random ssDNA library is an estimate of the minimum value, limited by the highest available VEGF concentration. K_d values of pool 3 (by FP and ACE) and pool 6 (by ACE) were not measured due to insufficient quantity of ssDNA produced in these rounds of selection.

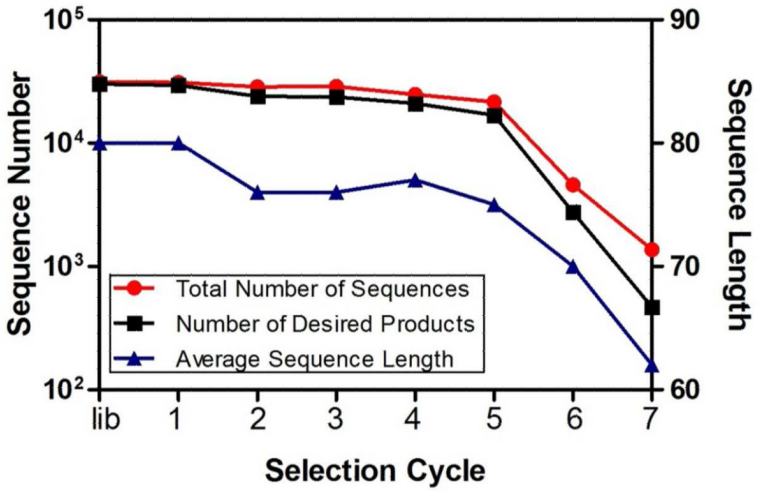

Figure 2.
Summary of sequencing results from each round of CE-SELEX selection comparing total number of sequences (●), number of desired products (■), and average sequence length (▲). Sequences with the lengths of 80 ± 2 nt are considered to be desired products.

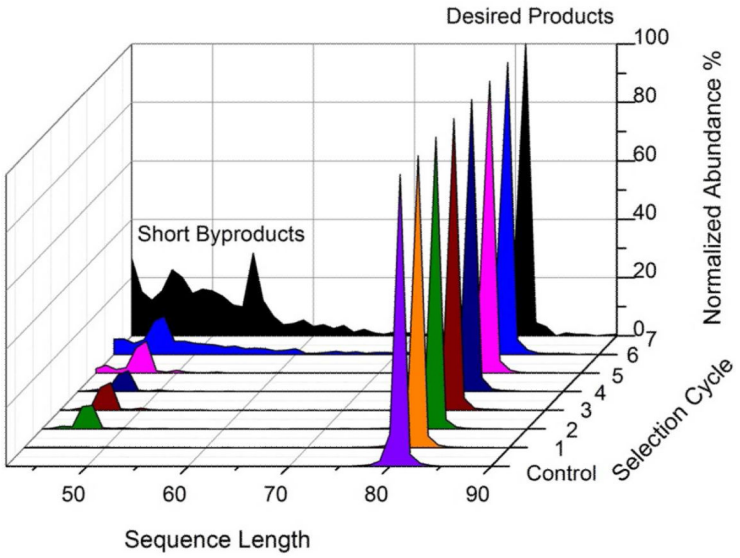

Figure 3.

Normalized sequence length distributions of the control and each selected pool. Each distribution was normalized to the abundance of the 80 nt sequences. Only sequences with the length equal to or above 42 nt were exported after 454 sequencing, so 42 nt was chosen as one boundary. 90 nt was chosen as the upper boundary since very few sequences were observed beyond this length

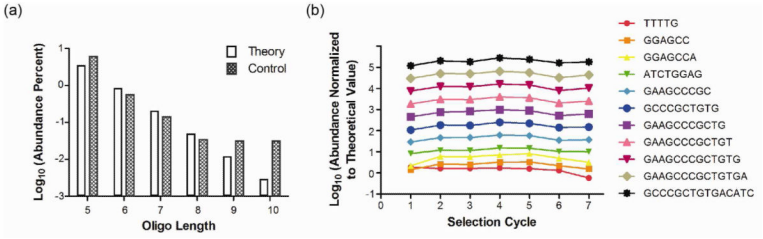

Figure 4.

(a) Comparison between the theoretical frequencies and the occurrences of the most abundant oligonucleotides in the control sample. (b) Evolution profiles of the most abundant 5 nt through 15 nt oligonucleotides.

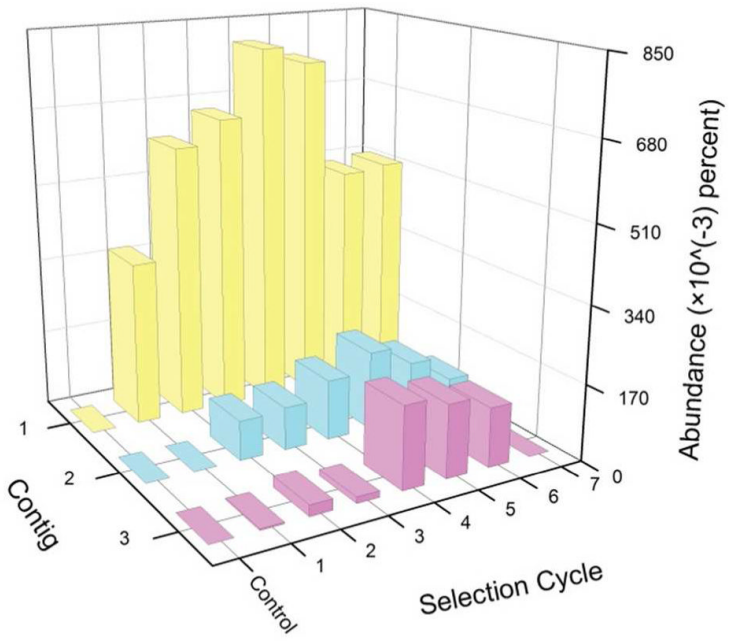

Figure 5.
Evolution profiles of the three most abundant aptamer contigs. The abundances ($\times 10^{-3}$ percent) of the top three contigs were shown over the 7 rounds of selection and compared with the control sample. Yellow: C 1, Blue: C 2, Pink: C 3.

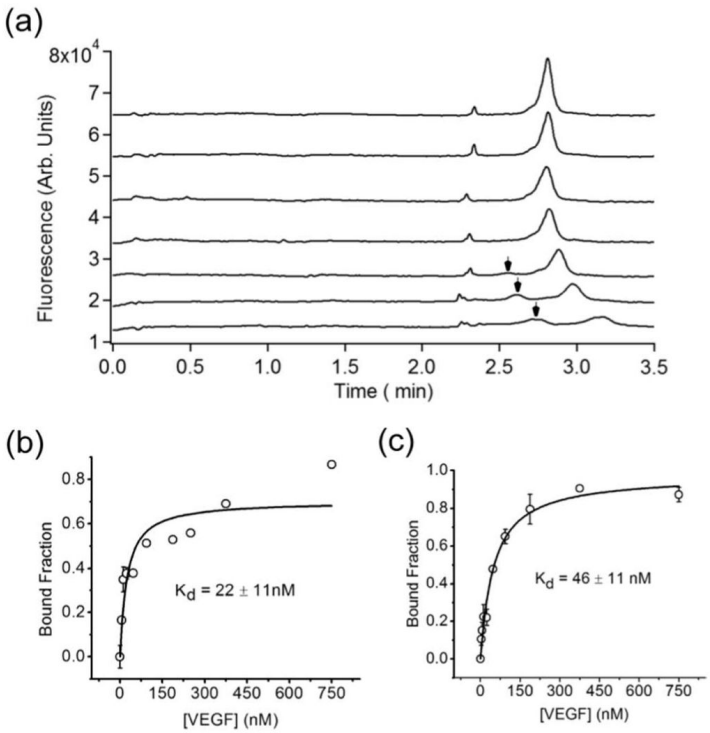

Figure 6.

(a) An overlay of electropherograms from one ACE experiment, titrating 5 nM R 3 with increasing concentrations of VEGF. The VEGF concentrations used were 0 nM, 5.9 nM, 12 nM, 23 nM, 94 nM, 375 nM, and 750 nM from the top to the bottom. The major peak was the unbound aptamer, and the minor peak indicated by the arrow was the VEGF-R3 complex. (b) The binding curve obtained from ACE experiments. (c) The binding curve obtained from FP experiments performed using the same samples. In (b) and (c) error bars report the standard deviation of three replicate measurements, and the errors of K_d represent the 95% confidence level.

Table 1

Analysis of the most abundant short sequences

		Initial library (control)	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	Pool 7
Length (nt)		45	50	47	47	47	47	47	54, 46, 42
Normalized abundance		0.157%	0.27%	8.08%	9.43%	7.28%	11.01%	13.13%	28.44%, 22.68%, 26.52%
Motif		CGTACTAC	CAGGTAGGGT	GGAGCCA	GGAGCCA	GGAGCCA	GGAGCCA	GGAGCCA	CGTGCTACCGTGAA, GAGCCA,
Normalized abundance of motif	Perfect match	0.004%	0.13%	5.04%	5.93%	4.82%	7.71%	6.57%	14.06%, 6.71%
	1 mismatch	0.024%	0.17%	6.75%	8.98%	6.95%	10.15%	8.40%	16.93%,9.27%
	MEME*	0.035%	0.18%	6.79%	9.18%	7.02%	10.49%	8.61%	20.45%,9.27%

* The p-value is equal to or smaller than 0.05 in MEME.

Table 2

Contigs (without primer regions) found in selected pools in order of descending abundance

Contigs (C #) and Randomly Picked Aptamers (R #)	Sequence (5′ to 3′)	length (nt)	Abundance		
			Total number	Number in Pool 4	$\times 10^{-2}$% in Pool 4
C 1	GGAAGCCCGCTGTGACATCTGGAGCCGGTTCCCGAGCCA	40	869	199	80.3
C 2	CGAGGATATCCGATTGTGGGTGACGTATTGGACACCTTCCCG	42	130	33	13.3
C 3	TGGCTGTTATTTCTCTCGGCCTCAGATCTTGGGCCCAACA	40	93	45	18.2
C 4	ATCCGGTTAGACCACTTGTCAATGAATGAGATTGTCAACG	40	37	8	3.23
C 5	TCTAGCGTTGTCTCGATCTATACCGTCGCGAATCGTAATG	40	29	6	2.42
C 6	TGCCATTCCTGTTGAGTATGACGTAATGGTAGTTGCGAAG	40	23	8	3.23
C 7	CGTTACCTGTAGGCGTATCTTTGTACAACCCTCTGGCGAG	40	17	7	2.82
C 8	AGTTTCTTGGGTTCTAAACGGATTTTATAAGGATTTTTAA	40	13	4	1.61
C 9	AAACTCAAAGACACTTTGGTTGAGTTGTGTGATACCCGTAG	41	11	0	0
R 1	GTCTGTACCAATCTAAAAGTGTATATTATATCTACGAGCCT	41	1	1	0.403
R 2	ATTCTGGGGAGTGGTACGTAAGCAGTGCCACCACGGTTAT	40	1	1	0.403
R 3	CCTTGTCATGATAGCTATGTTTCAGCCAACCATACCCGCT	40	1	1	0.403

Table 3

K_d values of selected aptamers for VEGF as measured by ACE and FP

Sequence	ACE*	FP*
C 1	49 ± 26 nM	101 ± 50 nM
C 2	21 ± 10 nM	162 ± 82 nM
C 3	66 ± 35 nM	164 ± 78 nM
R 1	53 ± 27 nM	100 ± 84 nM
R 2	72 ± 40 nM	88 ± 23 nM
R 3	22 ± 11 nM	46 ± 11 nM

* Errors represent 95% confidence intervals.