The in vitro selection world

Kenan Jijakli a,1, Basel Khraiwesh a,1, Weiqi Fu a,1, Liming Luo b, Amnah Alzahmi a, Joseph Koussa a, Amphun Chaiboontchoe a, Serdal Kirmizialtinc b, Laising Yen b, Kourosh Salehi-Ashtiani a,⇑

Laboratory of Algal, Systems, and Synthetic Biology, Division of Science and Math, and Center for Genomics and Systems Biology (CGSB), New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, United Arab Emirates

Department of Pathology & Immunology, Department of Molecular and Cellular Biology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA

Chemistry Program, Division of Science and Math, New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, United Arab Emirates

Article info

Article history:
Received 29 February 2016
Received in revised form 23 May 2016
Accepted 7 June 2016
Available online xxxx

Keywords:
In vitro selection
SELEX
RNA ligands
Aptamer
Next generation sequencing
Phage display
In vivo
In silico

Abstract

Through iterative cycles of selection, amplification, and mutagenesis, in vitro selection provides the ability to isolate molecules of desired properties and function from large pools (libraries) of random molecules with as many as 10^16 distinct species. This review, in recognition of a quarter of century of scientific discoveries made through in vitro selection, starts with a brief overview of the method and its history. It further covers recent developments in in vitro selection with a focus on tools that enhance the capabilities of in vitro selection and its expansion from being purely a nucleic acids selection to that of polypeptides and proteins. In addition, we cover how next generation sequencing and modern biological computational tools are being used to complement in vitro selection experiments. On the very least, sequencing and computational tools can translate the large volume of information associated with in vitro selection experiments to manageable, analyzable, and exploitable information. Finally, in vivo selection is briefly compared and contrasted to in vitro selection to highlight the unique capabilities of each method.

© 2016 Published by Elsevier Inc.

Contents

1. Introduction .......................................................... 00
2. A brief history of in vitro selection .................................. 00
3. In vitro selection of proteins ........................................ 00
4. Next generation sequencing and in vitro selection ............... 00
5. Computational tools and in vitro selection ......................... 00
6. Moving into the cell: In vivo selection .......................... 00
7. Selection beyond biology ........................................ 00
8. Conclusion and future perspectives .............................. 00
Acknowledgments ......................................................... 00
References ................................................................. 00

1. Introduction

How do you find a very small needle in a very large haystack? This proverbial problem is often used in conjunction with accomplishing a nearly impossible feat, and usually with a negative connotation to discourage attempts to such an endeavor. In vitro selection, a revolutionary approach developed in the 90s, solved this problem at a molecular level: This method enables isolation of functional molecules out of pools that are as large as 10^16 molecules. The methodology allowed expansion of the known catalytic and binding capabilities of DNA and RNA molecules, which at that time were not known to occur in nature. The approach has also been used to evolve and screen functionally improved proteins. The ability to isolate functional molecules from large pools of...
random molecules, often generated through combinatorial chemistry, is extremely powerful for finding extremely rare molecules. As such, in vitro selection may be described as a strategy with "irrationally designed" elements at its core [1] to denote the unbiased nature of the screens and importantly, the lack of need for a priori knowledge for obtaining new functional molecules.

The process of in vitro selection starts with a large pool of random molecules, followed by a functional screen. The pool enriched from the functional screen are then amplified and screened again. This series of screening, selection, then amplification is repeated with the pool of molecules many times, as low as four times and as high as twenty, until molecules with desired activity dominate the pool [2]. The power of this approach stems from its cyclic nature, which in turn introduces an exponential enrichment factor for isolation of the desired molecule. The cyclic nature of the experimental design can lead to numbers beyond astronomical scales: In an extreme case, investigators carrying out 300 rounds of selection through an innovative continuous selection scheme, were able to select the desired catalyst against a background of $3 \times 10^{238}$ fold amplification [3].

A typical in vitro selection experiment takes advantage of the ability to direct the chemical synthesis of large combinatorial libraries. Oligonucleotide pools containing many different sequences can be created with nearly any design. The basic scheme for in vitro selection experiments is outlined in Fig. 1. This process is referred to as SELEX (Systematic Evolution of Ligands by Exponential Enrichment) when it is applied specifically for the selection against ligands, with the obtained molecules called aptamers [4].

Although in vitro selection experiments are conceptually similar, the specific experimental designs may differ considerably based on the nature and characteristics of the target molecule and the selection strategy. In general terms, this translates to numerous experimental procedures each with their own efficiencies and limitations. Still, there are practical and inherent limitations that define the overall feasibility envelop for the selection success. For instance, the process generally requires an intermolecular interaction and information conservation; two requirements that limit the scope of possible molecular types in many in vitro selection experiments. For instance, a cis-linkage, covalent or otherwise, is often required between the active sequence and the product in order to select the active molecule. In addition, synthesizing and managing the starting molecular pools is defined by practical limitation of $\leq 10^{16}$ molecules for a typical laboratory setting experiment; this is not a theoretical limitation but a practical one. Another major limitation is that in vitro conditions often do not capture in vivo environments well, for instance, the cellular milieu of compounds, which typically includes thousands of different metabolites and proteins, reduces the translatability of in vitro results to in vivo environments as some of these variables may interfere with the function of the in vitro selected sequence. As such, the in vivo functionality of the molecules needs to be determined empirically once the selection is completed.

2. A brief history of in vitro selection

Numerous key discoveries in biology have emerged from studies of in vitro selection, including seminal work in the fields of biochemistry, genetics, molecular biology, molecular evolution and structural biology, in addition to therapeutic applications (Fig. 2).
Interestingly enough, the notion of in vitro selection was actually put forward nearly fifty years ago when Sol Spiegelman used a Qbeta replication system as a way to evolve a "self-replicating" nucleic acid molecule [29]. Two years later in 1971, Manfred Eigen presented a theoretical framework for the enrichment of molecular function due to selection under the context of molecular self-organization [30]. Later, Leslie Orgel built up on Spigelman’s work and introduced serial selection of ethidium bromide resistant V2 RNA [31]. Around 1989, all the techniques required for in vitro selection were starting to mature, which led Gerald Joyce to propose "in vitro evolution" as a method to probe the catalytic abilities of RNA [32]. In 1990, three labs independently developed the technique of selection: the laboratory of Jack Szostak at Massachusetts General Hospital and Harvard Medical School, coining the term in vitro selection for selecting RNA ligands against various organic dyes [4]; the laboratory of Larry Gold at University of Colorado, using the term SELEX for their process of selecting RNA ligands against T4 DNA polymerase [6]; and the laboratory of Gerald Joyce at the Scripps Research Institute, using in vitro selection to select for variant forms of a Tetrahymena ribozyme [7]. Jack Szostak and Andrew Ellington coined the term aptamer (from the Latin, apto, meaning ‘to fit’) for these nucleic acid-based ligands. Two years later, the Szostak laboratory and Gilead Sciences, independent of one another, used in vitro selection schemes to evolve single stranded DNA ligands for organic dyes and thrombin, respectively [8,9]. Also in the early to mid 90s, in vitro selection showed that DNA is capable of more complex folding mechanisms, and surprisingly that DNA can perform catalytic functions [8,12]. During the same period, new catalytic capabilities of RNA oligonucleotides were discovered or evolved using in vitro selection. Significant examples include the discovery of ribozymes that can ligate RNA molecules [10], and ribozymes that have kinase activity [33]. In 1994, in vitro selection moved beyond oligonucleotide selection and the first successful attempt to select proteins from random pools was reported through a strategy named “polysome display” [11]. The mid to late 90s saw the notion of continuous artificial evolution being introduced through uninterrupted in vitro selection experiments [3], and in the same period, other, more powerful protein in vitro selection approaches were introduced [13,14].

Many researchers have used in vitro selection as a means for application and discovery. In 2001, J. Colin Cox automated the process of in vitro selection in the Ellington lab, reducing the duration of a selection experiment from six weeks to three days [34–36], and at the same time in vitro selection experiments suggested multiple origins for the hammerhead ribozyme by selecting for similar motifs [16]. Shortly afterward, in vitro selection was used to produce artificial RNA based regulators termed riboswitches [37,38]. Later in 2006, Salehi-Ashtiani et al. [22] performed a genome-wide search for ribozymes and they found an HDV-like sequence in the human CPEB3 gene through an in vitro selection [22]. This marked the first attempt of combining in vitro selection approaches to whole genome analysis. In 2009, a novel class of chemically derived nucleic acids, called xeno nucleic acids (XNAs) was reported [39]. This was similar to the chemically derived threose nucleic acids that were reported in 2000 [40]. Both are oligonucleotides with an alternative sugar backbone that were shown to possess similar features to DNA and RNA, and aptamers from both XNAs and TNAs were developed using in vitro selection [26,41].

Please cite this article in press as: K. Jijakli et al., Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.06.003
3. **In vitro selection of proteins**

Recently, numerous studies on the development of *in vitro* selection tools and approaches have made tremendous progress in creating new DNA and RNA aptamers that potentially serve as novel oligonucleotide biomarkers. However, proteins and peptides serve wide and diverse functions owing to the diversity of functional groups in their side chains, and they are much more widely used in diagnostic, therapeutic, and industrial applications. Therefore, it would be of great interest to explore the extent of existing and new protein and peptide activities through developing *in vitro* selection methods for peptides and proteins with novel functions [17]. However, *in vitro* selection of proteins and peptides is more difficult than the selection of single stranded DNA and RNA molecules in principle and in practice, since encoding information for a protein sequence is usually missing after its translation step. Nevertheless, strategies and approaches for *in vitro* selection of high-affinity protein-binding peptides have been developed. Three typical approaches for *in vitro* protein selections are summarized in Fig. 3. The pros and cons of each approach are also summarized in Table 1. The full potential of *in vitro* protein selection technologies can be explored with the development of novel encoding strategies for libraries as well as the next-generation high-throughput sequencing technology. In addition, more efficient analytical tools.

Fig. 3. A schematic diagram of three different *in vitro* selection strategies for selection of peptides and proteins. Selection starts from construction of an initial DNA library. (a) Ribosome display. This strategy requires the formation of mRNA-ribosome-peptide complexes for binding and selection. (b) In mRNA display, a covalent bond between mRNA and the peptide is formed using puromycin as a linker. (c) SNAP display. It requires linking BG-labeled DNA to protein SNAP-tag fusion in water-in-oil droplets. BG, benzylguanine; SNAP-tag, the name for the DNA repair protein AGT (O6-alkylguanine DNA alkyltransferase), which was commercialized by Covalys AG in 2007 and distributed by New England Biolabs since 2009.
for functional characterization of peptides and proteins in medicine may accelerate the progress in in vitro selection of proteins.

Technologies for in vitro selection of peptides and proteins were first developed in the beginning of 1990s [11]. One such technology is phage display, which was developed as a simple and rapid in vitro selection approach for screening peptide libraries displayed (i.e. expressed) on the outer surface of phage particles. The premise is that the gene encoding the protein of interest is expressed as a fusion with a coating protein and, thus, is incorporated into the outside coating of a bacteriophage, which is genetically fused to a spacer sequence that lacks a stop codon. When translated, the spacer sequence remains attached to the peptidyl tRNA and occupies the ribosome tunnel, while the protein coding sequence protrudes out of the ribosome and folds.

Another novel approach was developed for protein in vitro selection by fusion puromycin, a peptidyl acceptor antibiotic that terminates translation, to a synthetic mRNA at its 3’ end to achieve stable mRNA-peptide fusions [14]. Whereas mRNA display, this approach was used to create a library of approximately 10^{12} random peptides and identify high-affinity aptamers to streptavidin, with up to four orders of magnitude higher affinity than those previously obtained by the phage display system, and with binding constants as low as 5 nM [17]. The same mRNA display strategy was also successfully applied to select ATP-binding proteins and yielded four new functional proteins from a library of 6 × 10^{12} proteins [44]. The mRNA display technology for in vitro selection can be a powerful approach for protein engineering and directed evolution by selecting peptides and proteins with desirable and functional properties from large libraries that consist of more than 10^{12} different protein mutants [45]. A similar concept was developed by Baskerville and Bartel who combined in vitro selection with rational design to produce a new ribozyme that forms a strong phosphoanhydride bond connecting RNA to protein [46].

Recently, the SNAP display system was developed as a protein in vitro selection tool by coupling genotype with phenotype via the SNAP-tag, which covalently binds DNA isolated in water-in-oil droplets to protein. The SNAP-tag is AGT ((O-6)-alkylguanine DNA alkyltransferase) expressed as a fusion to the studied protein, and forms a covalent thioester bond between the protein-coding DNA (bearing a covalently-linked SNAP-substrate, benzylguanine) and the expressed fusion protein. This results in increased stability and efficient selection of binding affinities ranging from nano to picomolar [47,48].

### 4. Next generation sequencing and in vitro selection

Next Generation Sequencing (NGS) has become a relatively cheap and user-friendly method, which can be applied with ease [49]. NGS has contributed substantially to in vitro selection techniques as in many other research fields including personalized medicine [50], molecular evolution, resequencing experiments [51] and cancer research [52]. In addition, NGS enables the investigation of in vitro selection pools [53–56] by revealing their sequence information, and relating the sequence information to function. NGS can also be used as a tool to study the evolution of ribozymes and other functional oligonucleotides. For instance, NGS can be used to track the change in ribozyme populations obtained through the in vitro selection of random sequences [57].

In recently developed in vitro selection techniques, NGS is used after each selection round instead of being applied after the last selection round as in previous applications of sequencing to in vitro selection [58–60]. NGS also enables comprehensive characterization of obtained aptamers, identification of their functional and rare motifs, and a comparison of functional motifs in each oligonucleotide population along with quantification of their abundance. NGS also allows for the study of the genetic evolutionary adaptation of oligonucleotide populations as the selection experiment proceeds. Pitt and Ferré-D’Amaré constructed an empirical fitness landscape for the evolution of a catalytic RNA by using NGS and a pool of 10^7 molecules obtained through in vitro selection [61].

In vitro selection in conjunction with NGS was first applied for the identification of CTF/NFI transcription factor (TF) ligands in genomic DNA [19]. Such methods, called SELEX-SAGE, combine SELEX with a serial analysis of gene expression. An additional
improved SELEX-SAGE method was used to identify binding specificities for 14 different classes of TFs [62]. Later, a combination of in vitro selection and NGS evolved into an experimental and computational platform called SELEX-seq, which was used to determine relative affinity of a DNA sequence to TFs [63]. This approach was used to characterize the binding specificity of eight Drosophila Hox proteins to DNA. An electrophoretic mobility shift assay was applied during three rounds of selection. NGS was performed after each selection round and computational analysis of sequencing data was applied to estimate the binding specificity of selected sequences [63]. SELEX-seq was also modified and employed for the identification of binding region of ESPR1 (epithelial splicing regulatory protein 1) [58]. A DNA library was transcribed by T7 polymerase into an RNA library, which was then incubated with a target and the bound sequences were reverse transcribed, amplified by PCR into cDNA and transcribed back into the RNA pool. The products were subjected to selection and single-end Illumina sequencing [58]. NGS has also been combined with electrophoretic selection techniques, namely capillary electrophoresis for the selection of thrombin binding aptamers, and it has been shown that a single round of capillary electrophoresis selection can enrich a random synthetic DNA oligo mixture for thrombin-binding activity from 0.4% aptamer content before selection to >15% aptamer content [64].

NGS is also advantageous for the selection of a large number of oligonucleotides. This approach was employed in the selection of genomic RNA aptamers recognizing various targets [65–67]. The process of enrichment of these aptamers was called genomic SELEX. It differs from the other types of SELEX by the starting pool of nucleotides it uses. A genomic DNA library is used in genomic SELEX in contrast to chemically synthesized oligonucleotide libraries that are used in classic in vitro selection [65,66]. Lorenz et al. [66] used genomic SELEX to map sequences in the E. coli genome and identify novel non-coding RNA aptamers binding to host factor Hfq (HF-I protein) of E. coli. This study has shown that genomic SELEX is a valuable technique in studying different regulatory domains of RNA.

Alternatively, total RNA pool from HeLa cells was used instead of genomic DNA for preforming transcriptomic SELEX [68]. It was reversely transcribed into cDNA, amplified by PCR and transcribed back into RNA by T7 RNA polymerase and HEXIM1 protein (hexamethylene bis-acetamide inducible protein 1) binding RNA. Cad mRNA, was identified using this approach [68]. Another method has been developed to analyze full sets of protein interactions using a combination of 454 pyrosequencing and an in vitro virus mRNA display method [69]. Specifically, this method covalently links proteins of interest to the mRNAs encoding them, and then detects the mRNA pieces using reverse transcription PCRs. The mRNA may then be amplified and sequenced. The combined method was titled IVV-HiTSeq and can be performed under cell-free conditions, though its results may not be representative of in vivo conditions [69]. IVV-HiTSeq can be used to generate high affinity protein interactions (protein–protein interactions, protein–RNA/DNA interactions and protein–chemical compound interactions). The IVV together with the NGS can elucidate the role of a particular protein in the pathway and biology of complex disorder diseases, for example cancer, by using integrone analysis (interactome with omics data) [70].

More recently, NGS has provided a means for aptamer functional selection in in vitro selection experiments. Nucleic acid aptamers were used to select for specific cell lines, and the abundance of nucleic acids on the cell surface was taken as a measure of functional efficacy. This abundance was measured using NGS which also identified the sequence identity of those aptamers [71].

In vitro selection together with the in silico analysis of next-generation high-throughput sequencing technologies, promises a better analysis of in vitro selection and improvements in the identification of aptamers.

5. Computational tools and in vitro selection

Various computational tools can be used in conjunction to in vitro selection experiments. Given the ease of access to such tools, and the significant advantage they possess, their usage is being more widely adapted. Given the large pool of molecules that is handled in in vitro experiments, there is no shortage of data to be handled and analyzed, but some computational tools perform other tasks that can greatly enhance the outcomes of in vitro experiments. Although beginning with a random pool can sometimes yield surprising and novel sequences, the efficacy of in vitro selection in yielding molecules with desired properties can be significantly increased by adopting computational tools in the process of selection. This is especially significant when selecting for a given catalytic function, since function is associated with structure and catalysis requires a relatively complex structure. The fully random pools that are used in selection are largely structurally simple. This is primarily because the molecules comprising those pools are constrained to smaller sizes [72,73]. Computational tools can therefore be used to ensure the incorporation of desired structural features in the starting random pool used in in vitro selection.

Specifically, computational tools that predict the secondary, or higher order structure of the proposed starting random pool can be particularly useful in creating pools customized for function. mFold, DNA Software’s OMP and ViennaRNA are examples of such tools, and all can be used to predict the secondary structure of oligonucleotides [74,75]. The central process that those pipelines rely on is the prediction of structure and from that associated function. Another tool that predicts two-dimensional (2D) and subsequently three-dimensional (3D) structures is MC-Fold/ MC-Sym [76]. Like most structural prediction tools, MC-Fold/ MC-Sym generates a set of 2D structures that correspond to a single sequence, and ranks those based on their minimal free energy, a widely accepted and benchmarked criteria [77–79]. A 3D simulation can then be performed on the selected 2D structures to generate a list of respective 3D structures.

Those tools can be incorporated in a systematic way into pool design such as to enrich for structural complexity and create a starting customized pool. Luo et al. [73] developed the Random Filtering algorithm for achieving that precisely. The Random Filtering algorithm starts with a random pool of RNA or DNA then uses structural prediction software to predict the secondary structure. Molecules in the pool are then scored based on a desired structural complexity, and the sequences more closely matching the desired complexity are selected and mutated. The mutations are random but cover every position in the sequence excluding the primer sites. This structural prediction, scoring, and then mutating is repeated until a pool of RNA or DNA molecules, enriched for structural complexity emerges [73].

Luo et al. [73] also proposed an algorithm for generating distributions of different structures as the starting pool for in vitro selection. Termed Genetic Filtration, the algorithm generates a number of pool designs with fixed primers, and each generated pool is scored based on how closely its structural distribution matches the desired distribution for the starting selection pool. Elements of the pool are then mutated, copied, and crossed until a pool with the desired distribution is reached [73].

Another alternative is to insert patterns within random sequences of the molecules in the initial pool. Those patterns are designed such as to ensure base pairing and formation of structurally more complex molecules [80]. Alternatively, Kim et al. [72] used mixing matrices and graph theory to design pools with
a desired structural distribution. Mixing matrices specify the proportion of different nucleotides in a given sequence, while graph theory was used to assess structural diversity and suggest motifs. A given weighted combination of a set of mixing matrices yields specific structural distributions, and this set and the respective weights are found iteratively. In all cases, designing for structural complexity seems to consistently lead to a richer pool of functional molecules [72].

Whether using designed pools or completely random pools, in vitro selection handles a large number of sequences, and sequence data organization and visualization can be insightful for in vitro selection experiments. Invis is a software aimed at that task exactly and it introduces visualization and organization tools that allow the user to examine the whole selection pool or individual sequences. Through an interactive approach, Invis also enables four primary analysis tasks; namely it can separate functional and non-functional sequences, identify recurring mutations and sequences, subdivide the functional sequences population, and reconstruct evolutionary trajectories between sequences [81]. There is a growing use of the high-throughput (HT)-SELEX (a combination of SELEX with sequencing technologies) and that requires the development of methods to gain insight into the selection process that combines the sequencing information. One recent tool that does that is AptaTools [82,83]. AptaTools is an automated and comprehensive pipeline for analyzing HT-SELEX data that performs data preprocessing, sequence analysis, cluster extraction, and data visualization. It also includes add-on tools such as AptaCluster, which clusters aptamer pools, AptaGUI, a graphical interface for visualizing AptaCluster results, and AptaSim, a program for simulation the selection process.

Yet another interesting consideration when conducting in vitro selection experiments is whether the isolated molecules have naturally occurring analogs or not, and here again, computational tools can be of assistance. Of course one quick and easy way to access tools for finding natural homologs is to use search databases and search algorithms such as BLAST (Basic Local Alignment Search Tool) and its varieties [84,85]; however, these tools frequently fail to identify functional RNA sequences as structural conservation (e.g., double-stranded stems) need not be accompanied by sequence conservation. To address this, some computational tools aim to find natural homologs through secondary structure predictions from genomic data. Examples include RNABob and RNAMST, both which take an RNA oligonucleotide secondary structure and an organism’s genome as inputs, and attempt to predict natural homologs through a combination of search and structural prediction algorithms [86,87]. A good example of the utility of such computational tools is the isolation of adenosine aptamers using genomic SELEX. Those aptamers were first predicted by RNABob before they were isolated using genomic SELEX [88]. A more recent and more efficient search tool is RNArobo, and it allows more flexibility in describing the secondary structure to be queried for [89].

Still, perhaps the ultimate goal when it comes to using computational tools for selecting molecules is to design the molecules using computational tools and screen them using computational tools before turning into wet lab experiments. This requires an approach that accurately predicts structure, function, dynamics as well as interactions with ligands. For now there is no such software available at hand. Instead, one can use multiple different software packages to elucidate each aspect. Chushak et al. [24] proposed a pipeline that uses a series of software packages that can allow for in silico screening of RNA ligands. First, patterned libraries are generated to ensure structural complexity, and then ViennaRNA is used to predict the secondary structure of the different RNA molecules in the library. The Rosetta software package was then used to predict their tertiary structures. AutoDock and Dock programs later were used to rank the docking abilities of each three dimensional RNA aptamer structures to a target ligand. Experimentally, six aptamer-ligand complexes were tested, and those in silico selected ligands were among the top 5% of the selected structures [24]. A similar computational pipeline has been adapted for the in silico selection of STAT (signal transducers and activators of transcription) inhibitory ligands. In this approach, Surflex Dock 2.6 was used to iteratively screen a library of ligands based on binding affinity. The selected compounds where then visually inspected using PyMOL before being validated experimentally [90].

Yet another computational tool that can potentially complement in vitro selection pipelines is all atom Molecular Dynamics (MD) simulations. Even though MD timescale falls short in studying the complete folding process, and hence cannot be used to predict the three dimensional structure from sequence data, MD simulations can be used as a refinement to such computational structural predictions by giving a more accurate physics based description of the three dimensional topology in solution and in finite temperature. Studying the fluctuations and local dynamics around the native basin of attraction can provide valuable insight into the structure-function relationship of the in vitro selected molecules. One approach is to generate three dimensional structures using a knowledge based software and test their stability using brute force MD in microsecond time scales. Also a set of experimental/functional constraints, such as the RMSD of an active site in a ribozyme can be used to select the correct fold. Experimental data on the selected molecules can also feed into molecular dynamics simulations to reduce the cost of conformational sampling and those simulations can elucidate the function mechanisms and inter-molecular interactions that characterize the selected molecule [91]. MD simulations can also be used to study the effects of minor perturbations such as mutations on the structure and functional performance of the selected molecule after the perturbation [92]. The predictions can always be tested experimentally, and that eventually leads to a more rationally enhanced design of the selected molecule.

Ultimately though, the use of computational tools in in vitro selection perhaps culminates in effective tools that can identify novel functional sequences as an alternative approach to in vitro selection. This of course is a difficult task and will be limited to classes of molecules that their structural constraints are well understood, but rational design of molecules has been applied successfully on subset of molecules to a certain extent. Those applications rely on biological databases and published experimental results, along with structural and dynamical modeling tools. Of course, certain molecules are easier to design or to undergo functional alteration through a design process. These molecules usually have more predictable chemistries and standard characteristics that have been observed through experiments, and design criteria can be formulated based on observations from those experiments. This is especially true in the case of functional oligonucleotides. One example is functional RNA molecules, including ribozymes, and one design computational tool used is RNAiFold. The algorithm of RNAiFold determines the RNA sequences required for a given structure through minimum free energy calculations primarily. It also accepts other constraints, such specific sequences, sequence number constraints and nucleotide content, that might be required for a particular function. RNAiFold has been used to design hammerhead ribozymes that were experimentally verified using activity assays [93].

Those approaches still require experimental validation and optimization, so consistent and sole reliance on computational design tools is still elusive. The challenge of sole reliance on computational tools is further exacerbated by recent developments in selection technologies and tools that make performing in vitro selection a highly efficient task.
6. Moving into the cell: In vivo selection

While “in vitro selection” is a powerful method for developing novel molecules with desired properties, it is challenging to extrapolate the products evolved from a test tube and apply them back to living cells. The environment within which nucleic acid sequences function in living cells is necessarily more complex with many additional variables than those governing the function of the same molecules in test tube. In the cellular environment, many factors, which are absent in a test tube, could theoretically limit the function of molecules evolved outside of cells (through in vitro selection). These factors include the alternative secondary and tertiary structures of RNA allowed within the cellular environment due to different metal ion concentrations (especially Mg$^{++}$), the constant turnover and transient nature of RNA within cells, and the unwanted interference due to protein-RNA, DNA-RNA and RNA-RNA interactions that are unique to cells. While an in vitro selection may have the advantages of speed and ease, molecules identified in this manner may be ineffective when evaluated in vivo. Therefore, the choice between an in vitro selection versus a cell-based selection is ultimately a choice between simplicity/efficiency versus intended applicability [94]. It is therefore helpful to keep the limitations of in vitro environments in mind while doing in vitro selection experiments, and effort should be put into mimicking in vivo environments as much as possible.

Cell-based selection however has its limitations too. First, the complexity of a library is limited by the size of the vector library (plasmids or virus) that can be practically constructed and introduced to cells, as well as by the size of the cell population that can be practically processed during selection. These factors effectively limit the complexity of the library to a range that is several orders of magnitude lower than in vitro selection. Second, the stringency or pressure required for effective selection may be toxic to cells on the time-scale of the assay, and this compromises the selection efficiency. Lastly, cell-based assays require longer time frames and more elaborate experimental procedures. Nevertheless, these disadvantages are offset by one major advantage of cell-based selection, that is, the molecules selected through this process are guaranteed to work in the intended cellular environment.

In vivo selection starts with a library of candidate molecules (DNA, RNA, or peptides), then expression of encoded molecules into the relevant cells. This is then followed by one of numerous selection techniques, such as flow cytometry using fluorescence reporters, drug resistance assay, RNA degradation assay, or other techniques that can specifically enrich the desired species. The critical steps of in vivo selection are illustrated in Fig. 4 using the in vivo selection of novel RNA aptamers for controlling RNA splicing in mammalian cells as an example. The aptamer employed in this case is capable of specific binding to tetracycline [95]; this property was previously used to control RNA splicing and translation in yeast cells [96,97]. By incorporating a splice site into the tetracycline aptamer and randomizing the critical areas of the aptamer sequence, an expression plasmid library was constructed and transfected into the human 293T cell population. Two different RNA products were expected in the transfected cells (Fig. 4a). In the absence of tetracycline, RNA splicing is expected to occur efficiently, and the shorter spliced RNA is the desired product. In the presence of tetracycline, which is expected to inhibit RNA splicing by binding to the aptamer, the unspliced full-length RNA is the expected product. The selection procedure is illustrated in Fig. 4b, which involves positive selection in the presence of tetracycline, and negative selection in the absence of tetracycline, and amplification/recovery of the desired RNA product to provide an enriched pool for the next cycle of selection. These cycles are repeated with increasing selection pressure (modulated by the concentration of tetracycline) until those species with desired properties have evolved from the population. In this particular case, RT-PCR of desired bands was used as the main selection technique. The selected sequences amplified by RT-PCR were digested with restrictions enzymes and then ligated back to the mammalian expression plasmid for the next round of selection. Note that in the specific strategy described, individual RNA, not individual cells, are the basis for selection. This general strategy could in principle be adapted to identify RNA sequences with various properties.

A more recent development in in vivo protein selection allows high-throughput analysis of protein variants expressed in vivo through the μSCALE (microcapillary single-cell analysis and laser extraction) platform. The protein variants are expressed in yeast or bacteria, and those cells are segregated and isolated using an array of glass capillaries. Once isolated, the cells can be imaged and analyzed and then extracted from their respective microcapillary using a laser. Once extracted, the cells can be cultured and further assayed [98].

There are also multiple considerations to keep in mind when using in vivo selection techniques. One important consideration pertains to genetic transfection or transformation. It is critical to optimize the transfection conditions in order to increase the representation of the library expressed in the mammalian cell population. This includes maintaining the cells in optimal conditions, identifying the suitable cell density for transfection and efficient transfection reagents for the chosen cells, and ensuring plasmid DNA quality and quantity required for efficient transfection. All
these factors need to be determined empirically prior to the in vivo selection. Another important consideration is the selection stringency. In the early phases of in vivo selection when the desired sequences are not highly represented, low selection pressure could be used to ensure the recovery of rare sequences. After two or three cycles of selection, when the desired sequences are enriched, the selection would benefit from using a higher stringency condition. Finally, library complexity is yet another vital consideration. The library size of in vivo selection is typically limited to a range of $10^6$–$10^7$ due to physiological constraints. One way to overcome this limitation is to intentionally introduce additional mutations during RT-PCR and PCR during each cycle of selection. An error-prone polymerase can be used for this purpose. Some commercial kits, such as GeneMorph II Random Mutagenesis Kit (Agilent), can also be employed to generate further mutations in a controlled manner.

7. Selection beyond biology

The generalized value of in vitro selection techniques comes from their practical adaption of evolutionary concepts. Randomness, mutations, and selection pressures all play a central role in in vitro selection, and the idea of practically and experimentally mimicking evolution is not unique to biological selections. It has been utilized in different circumstances and for many applications, and its essence is perhaps most adeptly and formally captured by Evolutionary Algorithms. This liberation of evolutionary pressure from the realm of experimental biology to computational abstractions can lead to significant and diverse applications. These include vehicle routing [99], airline crew scheduling [100], automotive design [101], evolvable hardware [102], marketing [103], and telecommunications networking [104].

Evolutionary algorithms start by assigning a fitness score to individuals in a population to be screened. A fitness score generally reflects the selection criteria; for instance, it can be the binding strength of a ligand, or the computational speed of a program, or the drag created in a mechanical design. After this initial evaluation and assignment of a fitness score, the individuals in a population are bred and also subjected to mutations at the same time. Breeding and mutation refer to a series of context specific operations that can be performed on the initial population to produce a new population, but in essence it is a sort of mixing the individuals in the old population to produce a new population, and to a lesser extent, this mixing is accompanied by random mutations or alterations to the new population that is produced. The probability of an individual in the original population to be selected for breeding increases the higher the fitness score of that individual is. This means that the fitter the individual in the population, the more likely they will be used for generating the new population, but even relatively unfit individuals can still be represented in the new population, albeit to a lesser degree. This inclusion of less fit individuals maintains diversity and has been shown to help in hastening the selection process in many instances. It has also been shown that the mutation rate in successive selection steps must be kept low for a quicker convergence to a solution. High mutation rates can sometimes introduce too much randomness that the selection does not yield any helpful results. In all cases, the algorithm is set up to terminate once given criteria have been reached (for instance an minimum binding affinity).

8. Conclusion and future perspectives

In vitro selection has been very successful, and that is evident from the numerous applications it has been used for. On the theoretical side, in vitro selection has provided experimental clues to evolution of function from random molecules, and that has contributed to the establishment of the plausibility of the RNA world hypothesis. Such studies have also contributed to understanding how and why molecular function arises. However, perhaps one of the most important contributions of the method is that key discoveries made through in vitro selection can be viewed to synergize with those in other fields, particularly in relation to the expansion of the role that functional RNA plays in extant cells.

The utility of in vitro selection has been amplified in recent years by combining other tools and techniques to the method. This has lead to the use of evolutionary methods for the selection of a wide array of molecules. NGS combined with in vitro selection experiments has proven its extraordinary efficiency and applicability and enabled the identification of different transcription factor specific binding motifs in DNA [62], and selection of RNA sequences that bind HIV-RT [59]. One additional development that couples in vitro selection and NGS could be to use deep sequencing to search for extremely rare functional variants in random nucleic acid pools and determine the activities of many sequences in parallel by following their frequencies over time. These tasks require the development of new computational tools capable of classifying millions of reads into families of functional nucleic acid folders while taking into account sequencing errors and PCR amplification biases. Generally, this means the development of computational tools for in vitro selection experiments that combines sequence management and sequence clustering with functional and structural prediction, and are able to integrate all that information. With the development and combination of these new tools, one can envision a faster development in research utilizing in vitro selection and consequently an ever-increasing acceleration in discoveries.

Acknowledgments

This work was supported in part by NYU Abu Dhabi Research Institute fund (G1205-12051) and NYU Abu Dhabi Faculty Research Fund (AD060) to K.S.; A. S.K. was supported by NYU Abu Dhabi Faculty Research Fund (AD181); L.Y. was supported by NIH R01EB013584 and a Duncan Cancer Center Pilot grant.

References

1. S. Brenner, R.A. Lerner, Encoded combinatorial chemistry, Proc. Natl. Acad. Sci. 89 (1992) 5381–5383.
2. D.S. Wilson, J.W. Szostak, In vitro selection of functional nucleic acids, Annu. Rev. Biochem. 68 (1999) 611–647.
3. M.C. Wright, G.F. Joyce, Continuous in vitro evolution of catalytic function, Science 276 (1997) 614–617.
4. A.D. Ellington, J.W. Szostak, In vitro selection of RNA molecules that bind specific ligands, Nature 346 (1990) 818–822.
5. A.R. Oliphant, A.L. Nussbaum, K. Struhl, Cloning of random-sequence oligodeoxynucleotides, Gene 44 (1986) 177–183.
6. C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, Science 249 (1990) 505–510.
7. D.L. Robertson, G.F. Joyce, Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA, Nature 344 (1990) 467–468.
8. A.D. Ellington, J.W. Szostak, Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures, Nature 355 (1992) 850–852.
9. L.C. Bock, L.C. Griffin, J.A. Latham, E.H. Vermaas, J.J. Toole, Selection of single-stranded DNA molecules that bind and inhibit human thrombin, Nature 355 (1992) 564–566.
10. D.P. Bartel, J.W. Szostak, Isolation of new ribozymes from a large pool of random sequences, Science 261 (1993) 1411–1418.
11. L.C. Mattheakis, R.R. Bhatt, W.J. Dower, An in vitro polynucleotide display system for identifying ligands from very large peptide libraries, Proc. Natl. Acad. Sci. 91 (1994) 9022–9026.
12. R.R. Breaker, G.F. Joyce, A DNA enzyme that cleaves RNA, Chem. Biol. 1 (1994) 223–229.
13. A.P. Jozef Hanes, In vitro selection and evolution of functional proteins by using ribosome display, Proc. Natl. Acad. Sci. 94 (1997) 4937–4942.
14. R.W. Roberts, J.W. Szostak, RNA-peptide fusions for the in vitro selection of peptides and proteins, Proc. Natl. Acad. Sci. 94 (1997) 1229–1230.
K. Jijakli et al. / Methods xxx (2016) xxx–xxx

Please cite this article in press as: K. Jijakli et al., Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.06.003
