Computational and Structural Biotechnology Journal 14 (2016) 262–270

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Dry and wet approaches for genome-wide functional annotation of conventional and unconventional transcriptional activators

Elisabetta Levati, Sara Sartini, Simone Ottonello *[,1], Barbara Montanini [1]

Biochemistry and Molecular Biology Unit, Laboratory of Functional Genomics and Protein Engineering, Department of Life Sciences, University of Parma, Parco Area delle Scienze 23/A, 43124 Parma, Italy

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 21 June 2016
Accepted 23 June 2016
Available online 29 June 2016

Keywords:
Transcription factors
Moonlighting transcriptional activators
Transcriptional activator trap
Yeast/bacterial one-hybrid
Protein binding microarrays
HT-SELEX

ABSTRACT

Transcription factors (TFs) are master gene products that regulate gene expression in response to a variety of stimuli. They interact with DNA in a sequence-specific manner using a variety of DNA-binding domain (DBD) modules. This allows to properly position their second domain, called "effector domain", to directly or indirectly recruit positively or negatively acting co-regulators including chromatin modifiers, thus modulating preinitiation complex formation as well as transcription elongation. At variance with the DBDs, which are comprised of well-defined and easily recognizable DNA binding motifs, effector domains are usually much less conserved and thus considerably more difficult to predict. Also not so easy to identify are the DNA-binding sites of TFs, especially on a genome-wide basis and in the case of overlapping binding regions. Another emerging issue, with many potential regulatory implications, is that of so-called "moonlighting" transcription factors, i.e., proteins with an annotated function unrelated to transcription and lacking any recognizable DBD or effector domain, that play a role in gene regulation as their second job. Starting from bioinformatic and experimental high-throughput tools for an unbiased, genome-wide identification and functional characterization of TFs (especially transcriptional activators), we describe both established (and usually well affordable) as well as newly developed platforms for DNA-binding site identification. Selected combinations of these search tools, some of which rely on next-generation sequencing approaches, allow delineating the entire repertoire of TFs and unconventional regulators encoded by the any sequenced genome.

1. Introduction

Transcription factors (TFs) coordinate many important biological processes ranging from cell cycle progression, cellular differentiation and development to intracellular metabolism and environmental adaptation [1–4]. Several human diseases, including cancer, are caused by alteration of regulatory programs and TFs are overrepresented among oncogene products [5]. About one-third of human developmental disorders is attributed to dysfunctional TFs [6] and programmed variations in the activity and/or specificity of TFs have also been documented as a major source of phenotypic diversity and evolutionary adaptation in various organisms [7–9]. Indeed, an increased complexity of TF-dependent regulatory networks is considered as a major driver of the emergence of metazoan life [10–13].

A distinguishing feature of typical ("conventional") TFs, compared to other transcriptional regulatory proteins, is their ability to interact with DNA in a sequence-specific manner. In the vast majority of cases, DNA-binding is achieved by one, sometimes more, DNA-binding domains (DBDs) and TFs are classified into superclasses and families according to the structural relatedness of their DBDs [14]. This DBD-based classification allows grouping different TFs on a structural basis. However, since the different structural motifs associated to the DNA-binding domains likely arose independently, this DBD-based structural classification does not necessarily mirror phylogenetic classification. In some cases, the DNA-binding domain provides clues on TF function. For example, homeo-domain containing TFs are often associated with developmental processes, while interferon regulatory factor family DBDs (helix-turn-helix motif) are functionally linked with the immune response [15], and fungal GATA factors are typically involved in nitrogen metabolism [16]. There are also proteins that display sequence-specific DNA-binding activity without any recognizable ("standard") DBD [17–19] and many orphan DBD types are likely to be still discovered and structurally classified. In addition to the DBD itself, other regions can contribute to, and influence, DNA-binding activity; for example, DBD-flanking regions directly involved in TF dimerization and function (e.g. [20]).

* Corresponding author at: Department of Life Sciences, University of Parma, Parco Area delle Scienze 23/A, 43124 Parma, Italy.
E-mail addresses: simone.ottonello@unipr.it (S. Ottonello), barbara.montanini@unipr.it (B. Montanini).
[1] These authors contributed equally as senior authors.

http://dx.doi.org/10.1016/j.csbj.2016.06.004

So-called "effector domains" are the other essential components of TFs. They mediate gene activation or repression by promoting the formation of active or repressed chromatin states, by directly or indirectly recruiting positively or negatively acting co-regulators (co-activators and co-repressors), or by modulating preinitiation complex formation or productive transcription elongation. At variance with the DBDs, effector domains are much less conserved and thus considerably more difficult to identify simply on a sequence similarity basis.

As a prototypical characteristics of TFs, sequence-specific DNA-binding is the main and first feature that is commonly addressed while trying to characterize (or discover) a new TF. The DNA-binding specificity of a TF, i.e., its ability to discriminate between different sequence motifs, is only one of several factors that can contribute to determine the sites it actually binds in the genome. In fact, DNA-binding site occupancy can also be influenced by site accessibility in a chromatin context, by cooperation or competition with other sequence-specific DNA-binding proteins, and by interaction with histones and other architectural proteins and chromatin modifiers as well. Circumstantial evidence in favor of this added layer of complexity is represented by the fact that most eukaryotic TFs tend to recognize short and degenerate DNA sequence motifs, as opposed to the larger motifs preferred by prokaryotic TFs [21].

Characterization of intrinsic sequence binding preference (i.e., the one referring to a simplified and restricted TF-DNA interaction) ideally requires either in vitro or heterologous assay systems allowing overcoming potential confounding effects caused by other modulating or competing TFs. Recent technological advances have greatly increased the speed and reliability with which (semi)quantitative estimates of DNA-binding ability and specificity can be obtained. These include a range of methods — e.g., microarray-based approaches [19,22–29] as well as high-throughput (HT) sequencing-based approaches [30–35], microfluidics-based technologies [36] and cell-based selection systems, also coupled with HT sequencing [37–40] aimed at increasing the number of DNA sequences that can be interrogated in parallel (outlined in Table 1).

Another emerging issue is that of so-called "moonlighting" transcription factors, i.e., proteins with an official function unrelated to transcription that play a role in gene regulation as either activators or repressors, as their second job. Cases of moonlighting ("unconventional") TFs, which are usually impossible to predict and particularly difficult to identify, have been documented in a variety of organisms ranging from bacteria to humans. For example, metabolic enzymes that moonlight as transcription factors, specifically designated as "trigger enzymes" or "metabolism-related transcription factors", which include enzymes directly or indirectly involved in gene expression regulation, with different documented or purported roles such as DNA/RNA binding, modulatory interaction with selected transcription machinery components, co-activator/repressor function and chromatin remodeling [18,19,41–43].

Here we present a general overview of the approaches, including both well-established as well as newly developed high-tech strategies, currently utilized for the functional analysis of TFs, highlighting their advantages and potential limitations. Particular emphasis is placed on genome-scale experimental methods that are accessible even to non-highly specialized molecular biology laboratories. Untargeted methods, also suitable for the large-scale identification of unconventional transcription factors, i.e., putative TFs lacking any recognizable DBD, are also discussed.

2. Delineating the transcription factor repertoire at the genomic level

Following genome sequencing, the first step in the identification and functional characterization of the transcription factor repertoire of a newly sequenced organism is the classification of the entire TF catalog based on the presence of conserved DBDs. The potential involvement of individual TFs in specific cellular processes can also be investigated based on sequence similarity with previously characterized transcription factors. TF functional validation can be then pursued with the use of a genome-wide approach such as the transcriptional activator trap (TAT) assay, which relies on the heterologous expression of cDNA libraries or specific TF subsets in the yeast *Saccharomyces cerevisiae*. This method allows the rapid characterization of the transcriptional activation capabilities of predicted TFs. Because of its untargeted nature, the TAT assay also allows the identification of new putative unconventional activators lacking any recognizable DBD, which escape detection by search methods strictly based on sequence similarity.

2.1. TF identification and classification

Sequence-specific TFs are thought to comprise between 0.5 and 8% of the eukaryotic gene content and can be classified into superclasses and classes according to the structure of their DBDs [14,44]. DBDs display a wide range of structural motifs encompassing a diverse array of protein folds, each representing a different solution to the problem of sequence-specific DNA recognition. More than 80 and 60 different DBD types have been recognized to-date in eukaryotes and prokaryotes, respectively, with very few DBD types shared between these two

Table 1
Outline of in vitro and in vivo heterologous high-throughput DNA-binding assays[a].

Acronym	Name	Throughput	Probe type	Resolution[b]	References
HT-SELEX	High throughput systematic evolution of ligand by exponential enrichment	10^{15}	Oligo library	Qualitative (SELEX) Quantitative (HT-SELEX)	[30,32,34,104]
Bind-n-Seq	Bind and sequence	10^{13}	Oligo library	Quantitative	[35]
HiTS-FLIP	High throughput sequencing-fluorescent ligand interaction profiling	10^{9}	Oligo library (clusters on Illumina flow cell)	Kinetics	[31]
B1H	Bacterial one-hybrid	10^{8}	Oligo library (in plasmid)	Qualitative (B1H) Quantitative (B1H followed by HT-seq)	[37,39,84,87]
PBM	Protein-binding microarray	10^{6}	Microarray	Quantitative	[22,24–27,90,91,101,107]
CSI	Cognate site identifier	10^{6}	Microarray	Quantitative	[29,92,108]
EMSA-seq	EMSA followed by high throughput sequencing	10^{5}–10^{6}	Oligo library	Quantitative	[33]
MEGAshift	Microarray evaluation of genomic aptamers by shift	10^{3}	Oligo library	Quantitative	[28]
MITOMI	Mechanically induced trapping of molecular interactions	10^{2}–10^{3}	Oligo library	Kinetics	[36,93,94,109]
HT-SPR	High throughput surface plasmon resonance	10^{2}	Microarray	Kinetics	[110,111]
TIRF-PBM	Total internal reflectance fluorescence PBM	10^{2}	Microarray	Kinetics	[23,112]

[a] High-throughput, TF binding site discovery approaches ordered by throughput, i.e. the approximate number of DNA sequences interrogated in parallel (as reported in the cited references); probe type refers to the specific format of the DNA probe oligomer utilized by each method.
[b] Qualitative: only binding sites with the highest affinity are likely to be obtained; Quantitative: binding models (e.g. PWM) can be determined; kinetics: equilibrium binding specificities and kinetic constants can be calculated.