Nucleic Acids Research, 2025, **53**, gkaf140
https://doi.org/10.1093/nar/gkaf140
Synthetic Biology and Bioengineering

Decoding and reengineering the promoter specificity of T7-like RNA polymerases based on phage genome sequences

Jinwei Zhu [1,2], Ziming Liu [3], Chunbo Lou [3], Quan Chen [1,2,4,*], Haiyan Liu [2,4,*]

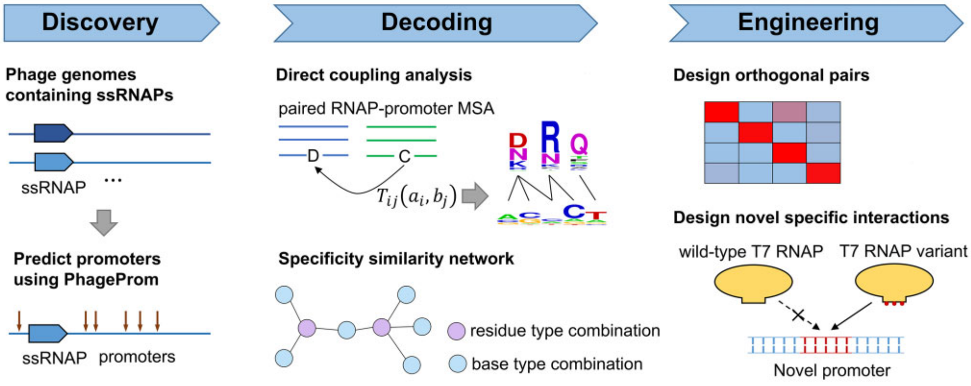
[1] Department of Rheumatology and Immunology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, Hefei National Research Center for Physical Sciences at the Microscale, University of Science and Technology of China, Hefei, Anhui 230001, China
[2] MOE Key Laboratory for Membraneless Organelles and Cellular Dynamics, School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230027, China
[3] Center for Cell and Gene Circuit Design, CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[4] Biomedical Sciences and Health Laboratory of Anhui Province, University of Science and Technology of China, Hefei, Anhui 230027, China

[*] To whom correspondence should be addressed. Email: hyliu@ustc.edu.cn
Correspondence may also be addressed to Quan Chen. Email: chenquan@ustc.edu.cn

Abstract

The single subunit RNA polymerases (ssRNAPs) of bacteriophages are highly interesting targets for the prediction and engineering of specific protein–DNA interactions. Despite extensive existing studies focusing on particular ssRNAPs such as the T7 RNAP, few rules governing the protein–DNA sequence covariations across diverse ssRNAPs and their cognate promoters are clearly known. Here, aiming to reveal such rules, we comprehensively mined promoters of various categories of ssRNAPs from phage genomes. For T7-like RNAPs, direct coupling analyses of the predicted set of RNAP–promoter pairs revealed that the interaction specificity was dominantly encoded by the amino acid and nucleotide residues at only a few key positions. The covariations between the amino acid and the nucleotide residues at these positions were summarized into a sparsely connected network. Using experimentally verified connections in this network, we designed a set of orthogonal T7 RNAP–promoter variants that showed more stringent orthogonality than previously reported sets. We further designed and experimentally verified variants with novel interactions. These results provided guidance for engineering novel RNAP–promoter pairs for synthetic biology or other applications. Our study also demonstrated the use of comprehensive genome mining in combination with sequence covariation analysis in the prediction and engineering of specific protein–DNA interactions.

Graphical abstract



Introduction

Single subunit DNA-dependent RNA polymerases (ssRNAPs) from bacteriophages [1], such as the T7 RNAP and its relatives (e.g. T3 and SP6), are highly specific for their individual promoters and widely used for RNA synthesis [2]. It is of substantial interest to discover how the promoter specificity of different members of this RNAP family depends on their amino acid sequences [3, 4], and to engineer new variants of these RNAPs to achieve non-native and/or orthogonal promoter specificity [5, 6].

Received: November 1, 2024. **Revised:** February 2, 2025. **Editorial Decision:** February 10, 2025. **Accepted:** February 14, 2025

Broadly speaking, these ssRNAPs perform their function through binding to the genome DNA, just like many other DNA-binding proteins. It is possible to discover the DNA binding specificity for such proteins through mining the sequenced genomes of various organisms [7–10]. This approach has been successfully applied to various protein families including transcription factors [11–13], nucleases [14], and recombinases [15]. It has also been applied to the phage ssRNAPs [16, 17]. Currently, the number of publicly available bacterial phage genomes exceeds 20000 [18]. Comprehensively mining these genomes can substantially increase the number of predicted cognate ssRNAP–promoter pairs. The resulting dataset may serve as a valuable resource for discovering and engineering diverse ssRNAP-based expression systems, enabling applications such as gene expression in industrial non-model organisms [17, 19] and RNA production with reduced byproducts [20].

Existing tools for predicting phage promoters associated with ssRNAPs include PHIRE [21] and PhagePromoter [22]. PHIRE predicts promoters by searching individual genomes for repetitively occurring DNA motifs (∼20 bp). PhagePromoter filters DNA motifs using sequence matching scores against several known "standard" promoter sequences (e.g. the promoter sequence 'TAATAAGACTCACTAAAGGGAGA' of the T7 RNAP). While setting up the basic framework for predicting phage promoters, neither PHIRE nor PhagePromoter is suitable for comprehensively mining thousands of phage genomes. For instance, PHIRE is prone to interference from non-promoter repetitive motifs in genome sequences, as evidenced by incorrect predictions in the genomes of *Yersinia* phages Yepe2 and Yep-phi and *Caulobacter* phage Percy [17]. Additionally, PHIRE takes up to several minutes to process a single genome, which makes it computationally inefficient for large-scale genome mining. As regards PhagePromoter, a major limitation is that reliable predictions are restricted to promoters that are highly similar to the "standard" promoters. Moreover, it relies on the correctness of the "standard" promoter sequences themselves.

As a result of the above drawbacks of existing tools, the available data sources of promoters of phage ssRNAPs are limited, covering only a small fraction of known phage genomes. One existing dataset containing 662 T7-like promoters (i.e. promoters of T7-like RNAPs) was derived from PHIRE predictions on only 46 genomes [17]. Another dataset, which was used as the training dataset for PhagePromoter, included only 135 T7-like promoters from 10 phage genomes out of 800 promoter sequences retrieved from the phiSITE [23] database and available publications for 53 phage genomes. Additionally, the "standard" promoters used by PhagePromoter for several other RNAPs, including phiKMV and KP34, do not agree with recently reported experimentally validated promoter sequences of these RNAPs [24, 25]. In general, the limited existing data preclude in-depth analyses of relevant protein–DNA sequence covariations.

A comprehensive and reliable dataset of natural phage RNAP–promoter pairs may help decode the promoter specificity of ssRNAPs through the analysis of protein–DNA sequence covariations. The results may facilitate the engineering of RNAPs with new promoter specificity for various applications. For example, T7 RNAP variants with engineered orthogonal promoter specificity have been employed for the modular control of complex gene expression in synthetic biology [26, 27] as well as for multiplexed monitoring of protein–protein interactions [28]. Previous attempts to create orthogonal T7 RNAP–promoter variant pairs relied on part mining and domain grafting [16], or directed evolution [6, 29]. For example, Temme *et al.* have used the RNAP–promoter pairs predicted with PHIRE [21] to guide domain grafting, where the promoter recognition loop in T7 RNAP was replaced with loops from RNAPs of T3, K1F or N4. The resulting sets of RNAP–promoter variants indeed displayed a limited extent of orthogonality, but non-cognate RNAP–promoter pairings still showed relatively high activity. Directed evolution could extend the promoter spectrum of an RNAP but could not easily achieve orthogonal promoter recognition without negative selection [30]. The phage-assisted continuous evolution (PACE) method incorporating negative selection [31] can effectively eliminate cross-pair activity. It was successfully employed to construct two pairs of highly orthogonal T7 RNAP–promoter variants. However, it is challenging to evolve more than two orthogonal pairs using negative selection.

In the current work, we carried out comprehensive mining of cognate ssRNAP–promoter pairs from publicly available phage genomes. Promoters for several categories of ssRNAPs, including T7-like, phiKMV-like, KP34-like, and Syn5-like RNAPs, were predicted. We analyzed the covariations of the protein–DNA sequences in the predicted set of T7-like RNAP–promoter pairs with a pseudolikelihood maximization direct coupling analysis (plmDCA) approach [32], and revealed that the specificity of the protein–DNA interaction of T7-like RNAPs was dominantly encoded by the amino acid residue types and base types at only several key positions. The relationships between the combinations of amino acid residue types and of DNA base types at these key positions were visualized as a sparsely connected network. After systematically testing the connections in this network by experiments, we successfully designed and experimentally verified T7 RNAP–promoter variant pairs with orthogonal specificity as well as with new specific interactions.

Materials and methods

Mining promoters of ssRNAPs from phage genomes

Our computational workflow for mining cognate ssRNAP–promoter pairs from phage genomes has been devised based on several key observations. First, with most phages lacking their own RNAP and instead using the host RNAP, the ssRNAPs used by a small number of phages are homologous to the T7 RNAP [33]. Second, these ssRNAPs are commonly grouped into different categories including the T7-like, SP6-like, phiKMV-like, and P60-like RNAPs based on their similarity to certain representative ssRNAP family members [24]. Third, a small number of promoters for ssRNAPs in different categories have been experimentally characterized, including promoters for T7, T3, SP6, LUZ19 (phiKMV-like), KP34 (phiKMV-like), and Syn5 (P60-like) RNAPs [24, 25, 34–37]. These promoters usually occur multiple times within a genome (T7 promoters appear 18 times, SP6 promoters appear 12 times, LUZ19 and KP34 promoters appear four times, and Syn5 promoters appear twice), with the promoter sequences within the same genome being highly conserved.

Our computational workflow, named PhageProm, is shown in Fig. 1A. The workflow started with the identification of phage genomes that contain ssRNAPs. To perform this, we downloaded 23854 phage genomes from the website http://millardlab.org/bacteriophage-genomics/phage-genomes-1Dec-2022/, built a local database using the 'makeblastdb' command of the ncbi-blast-2.13.0 + program [38], and searched the local database with the same program by using the amino acid sequence of T7 RNAP as a query (no E-value cut-off was specified). The results gave us a list of ssRNAP-containing phage genomes.

From these genomes, we subsequently predicted promoters for various categories of ssRNAPs in two steps. Step one was to search for repetitively appearing sequence motifs (each represented as a group of DNA subsequences) from each genome. Step two was to pool together the sequence motifs found in different genomes containing the same category of ssRNAPs and to identify promoters through sequence clustering.

To perform the first step on a given genome, we considered all contiguous segments in a genome sequence with a length of L = 20 bp (with L − 1 bp overlaps between two adjacent segments). Each segment was considered in turn as a Query subsequence. For a given Query, all 20 bp segments (Subject subsequences) separated by >100 bp from the Query were matched one-by-one against the Query. If the Hamming distance between a Subject subsequence and the Query was no more than 4 bp, the Subject subsequence was collected. A Query and all of its matched Subject subsequences comprised a Query:Subject group. To filter out groups that contained segments in repeat elements of the genome, we removed every group that contained three or more matched Subject subsequences that were within 200 bp in their genome locations. Then any two Query:Subject groups were merged into a single group if their matches corresponded to contiguous subsequences that overlapped by >10 bp. The merging was performed by extending both the Query and the Subject subsequences, and was performed repeatedly until no further merging was possible. The groups after merging were sorted according to the number of subsequences in descending order. Then, redundant groups were removed from the sorted list by aligning each of the subsequence in a group to the Query of every preceding group. If ≥ 10 bases were found to be identical in a gapless alignment, the group containing the smaller number of subsequences was removed from the list. This produced a final sorted list of non-redundant groups of Query:Subject for each genome, each group corresponding to a DNA sequence motif represented by one Query and one or more Subject subsequences that match the Query.

To carry out the second step, we pooled and clustered the Query:Subject subsequence groups from the phage genomes containing a specific category of ssRNAPs. To predict promoters for the T7-like, phiKMV-like, KP34-like, and Syn5-like RNAPs, we identified genomes containing ssRNAPs with similarity (ncbi-blast E-value $< e^{-50}$) to the respective reference proteins (i.e. T7 RNAP, phiKMV RNAP, KP34 RNAP, and Syn5 RNAP). We then pooled the Query:Subject groups from these genomes separately for each RNAP category. The minimum number of subsequences in each of the selected Query:Subject groups was required to be five for T7-like RNAPs, three for phiKMV-like RNAPs, and KP34-like RNAPs, and two for Syn5-like RNAPs. Notably, some genomes contained RNAPs that could be grouped to multiple categories based on the ncbi-blast searches. Subsequences from these genomes were added to the respective pools in parallel.

The subsequences in each pool were clustered using the DBSCAN algorithm [39], which requires three user-defined parameters: the distance function (*distFunc*), the cut-off radius (*eps*) for the neighborhood of a reference point in the sequence space, and the minimum number of points (*minPts*) in the sequence space required to form a dense region. Here, we defined the *distFunc* as $1 - \text{Match}(x, y)/Length(x)$, where $\text{Match}(x, y)$ refers to the number of identical bases between subsequences x and y, and $Length(x)$ denotes the total number of bases in subsequence x. The parameters *eps* and *minPts* were set to 0.2 and 2, respectively. Among the clusters of subsequences of various pools, we could identify clusters containing the known promoter sequences of the T7, LUZ19 (phiKMV-like), KP34, and Syn5 RNAPs, as well as clusters containing the reverse complementary sequences of these promoters. Subsequences within the former clusters were predicted as promoters (below referred to as T7-like promoters, phiKMV-like promoters, etc.) associated with the corresponding categories of ssRNAPs.

Constructing position-specific scoring matrices (PSSMs) from predicted promoters

The predicted promoters for each type of RNAP were aligned without gaps, and PSSMs of 21 bp in length were constructed by computing the frequency of each nucleotide type b at each position i as

$$f_i(b) = \frac{n_i(b)}{N}, \tag{1}$$

where $n_i(b)$ is the number of occurrences of nucleotide type b at position i, and N is the total number of promoter sequences. The log-odds score for each nucleotide type at each position relative to the corresponding background distribution was computed as

$$s_i(b) = \log_2\left(\frac{f_i(b)}{p(b)}\right), \tag{2}$$

where $p(b)$ is the background probability of nucleotide type b, which was set to 0.25.

Analyzing protein–DNA joint multiple sequence alignment

We applied plmDCA to analyze the multiple sequence alignment (MSA) of the predicted cognate RNAP–promoter sequence pairs for the T7-like RNAPs. The protein–DNA joint MSA can be represented as

$$\{(a_1^m, \ldots, a_L^m, b_1^m, \ldots, b_N^m)\}_{m=1,\ldots M}, \tag{3}$$

where M is the number of rows in the MSA, and L and N are the numbers of columns of the protein and DNA parts, respectively. $a = (a_1^m, \ldots, a_L^m)$ represents the aligned protein sequence and $b = (b_1^m, \ldots, b_N^m)$ represents the aligned DNA sequence.

We considered learning both bi-DCA and uni-DCA models from the protein–DNA joint MSA. In the bi-DCA model, the probability of a combined protein–DNA sequence is represented as

$$P(a,b)=\frac{1}{Z}exp\left[\sum_{1\le i\le L}h_i^a(a_i)+\sum_{1\le i\le N}h_i^b(b_i)+\sum_{1\le i<j\le L}J_{ij}^a(a_i,a_j)+\sum_{1\le i<j\le N}J_{ij}^b(b_i,b_j)+\sum_{1\le i\le L,1\le j\le N}T_{ij}^{ab}(a_i,b_j)\right]. \qquad (4)$$

The five energy terms in the above formula correspond to: the energies of individual protein residues, the energies of individual DNA bases, the couplings between protein residues, the couplings between DNA bases and the couplings between DNA bases and protein residues. In the uni-DCA model, the energies of individual protein residues and the couplings between protein residues are ignored, and the probability is defined as

$$P\left(b\right)=\frac{1}{Z}exp\left[\sum_{1\le i\le N}h_i^b\left(b_i\right)+\sum_{1\le i<j\le N}J_{ij}^b\left(b_i,b_j\right)+\sum_{1\le i\le N,1\le j\le L}T_{ij}^{ba}\left(b_i,a_j\right)\right]. \qquad (5)$$

Compared with bi-DCA, uni-DCA has a significantly reduced number of parameters, especially when the protein sequence is long. A negative control with uni-DCA was devised by randomly shuffling the protein and DNA sequences within each row of the joint MSA. More details for learning the parameters of the DCA models are given in Supplementary Methods.

Experimental assays of RNAP–promoter activity

Plasmid construction

We constructed a genetic circuit containing a pair of T7 RNAP–promoter variants of interest along with a green fluorescent protein (GFP) reporter (Supplementary Fig. S1). Since the mutations in the T7 RNAP variants (N748, R756, and Q758) were all located within Region3 (K740–E768), the Golden Gate Assembly method was employed to efficiently assemble all genetic fragments into plasmids for each pair of T7 RNAP–promoter variants. Primers A-F and A-R were designed to amplify the Region3 fragment, incorporating both the variant sequence and adapter required for assembly. Similarly, primers P-F and P-R were used to amplify each corresponding promoter fragment (Supplementary Fig. S2).

Growth and fluorescence measurement

The plasmids were transformed into 25 μl of Trans10 Chemically Competent Cell (TransGen Biotech). After recovery, the cells were cultured overnight at 37°C on LB agar plates containing 100 μg ml^{-1} ampicillin. Two colonies were picked, grown separately, and sequenced to confirm the correct Region3 and promoter sequences. The validated clone was inoculated into 0.9 ml of LB medium [supplemented with 100 μg ml^{-1} ampicillin and various concentrations of isopropyl-β-D-thiogalactopyranoside (IPTG)] in a 96-deep-well plate and cultured for 16 h at 37°C and 250 rpm. Specifically, the activities of 137 pairs in the specificity and similarity network (SpSN) were measured under 10 μM IPTG induction, while the activities of 16 orthogonal pairs and 33 pairs with novel specific interactions were measured under 50 μM IPTG induction. This process was repeated three times.

Subsequently, 2 μl of the cell culture was transferred to a 96-well plate containing 150 μl of phosphate-buffered saline supplemented with 1 mg ml^{-1} kanamycin for translation inhibition. Fluorescence was measured using a BECKMAN CytoFLEX S flow cytometer, with excitation at 488 nm and detection at the fluorescein isothiocyanate (FITC) channel (525/40 nm). Ten thousand events per sample were recorded, and the fluorescence intensity was calculated using CytExpert software. The activity of each T7 RNAP–promoter variant pair was determined as the average fluorescence from triplicate samples. Three internal control pairs, NRQ–CGACT (wild-type), DRQ–CCCCT, and DRQ–CGACT, were included, and their activity values were consistent with those reported in previous studies [6] (Supplementary Fig. S3).

Similar to previous studies [16, 40–43], we also encountered the cellular toxicity issue of the wild-type T7 RNAP–promoter pair in *Escherichia coli* cells. All the T7 RNAP–promoter variant pairs were successfully constructed, except for the wild-type pair. The wild-type pair experienced construction collapse, i.e. spontaneous mutation of T7 RNAP or the T7 promoter (Supplementary Fig. S4A). This indicated that T7 RNAP or the T7 promoter alone was not toxic, but their combination caused severe growth defects, possibly due to the excessively high transcriptional activity. To resolve this issue, we used a very weak ribosome-binding site (RBS) to reduce the production of T7 RNAP (Supplementary Fig. S4B). Therefore, this weak RBS was constructed into the genetic circuit (Supplementary Fig. S1) for measuring the activities of T7 RNAP–promoter variant pairs.

Results

Predicted ssRNAP promoters in phage genomes

A total of 1757 phage genomes containing ssRNAPs were identified with ncbi-blast. Subsequently, we identified 11 218 T7-like promoters in 840 genomes, 467 KP34-like promoters in 135 genomes, 186 phiKMV-like promoters in 62 genomes and 18 Syn5-like promoters in nine genomes. Category-specific PSSM scores were derived, and Fig. 1B shows the sequence logos for these four categories of promoters. For 741 genomes, no definite prediction of the promoters for their ssRNAPs was made. However, we provided a number of groups (on average 20 groups per genome) of repeated subsequences for each genome, each containing at least three subsequences. Additionally, we scanned each genome using the four category-specific promoter PSSM models shown in Fig. 1B, and selected for each PSSM the 20 highest scoring subsequences as potential promoters. We expect that these pre-selected, limited numbers of repetitively occurring or high-scoring subsequences can substantially narrow the search range for experimentally identifying promoters in any of the processed genomes. All predicted promoters, subsequence groups, and potential promoters are provided in Supplementary Dataset S1.

It is interesting to compare the predictions reported here with those from PHIRE and PhagePromoter, the two previous methods that inspired the development of PhageProm. A major improvement of PhageProm over PHIRE is that PhageProm implements a mechanism to avoid a major type of error of PHIRE: predicting genomic repeat elements as promoters

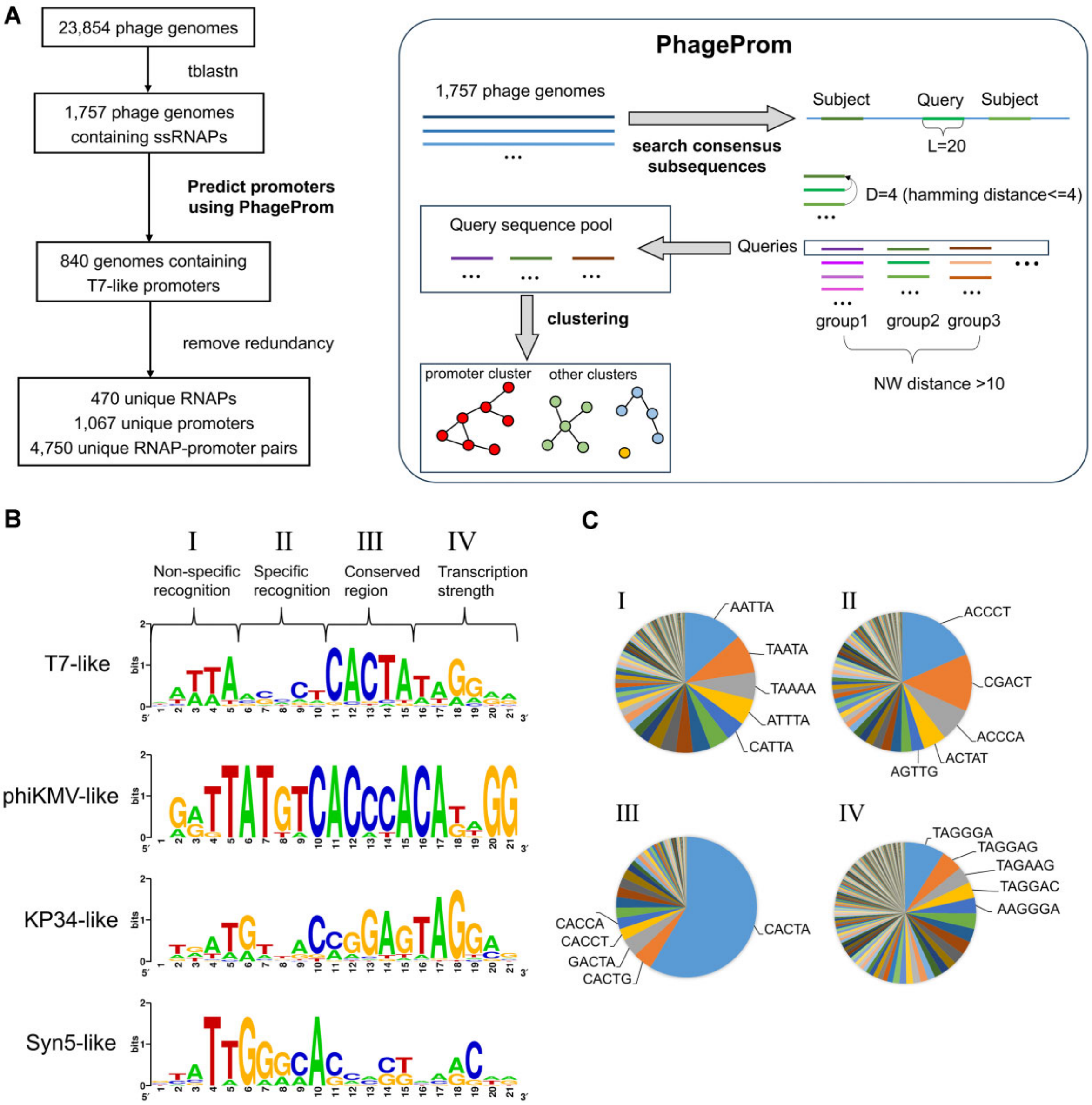


Figure 1. Mining and analysis of promoters of ssRNAPs from phage genomes. (**A**) An overview of the computational pipeline. (**B**) Sequence logos for promoters of four categories of ssRNAPs. For the T7-like promoters, the 21 positions are divided into four regions. (**C**) Pie charts displaying the frequency of exact nucleotide subsequences in each promoter region.

(Supplementary Table S1). Another improvement is that, unlike PHIRE, PhageProm does not make the assumption that the promoters must correspond to the most frequently occurring motif in a genome. This allowed PhageProm to discover phage promoters outside of the T7-like category, which may not occur so frequently in the genomes. Moreover, even the T7-like promoters are not always the most frequent motif in some phage genomes, causing PHIRE to generate false-positive predictions in these genomes (Supplementary Table S2). To illustrate this, we compared PhageProm and PHIRE on the 46 genomes from the dataset of Zhao *et al.* [17], which included a total of 662 T7-like promoters predicted by PHIRE. Although the PHIRE predictions for 40 out of 46 genomes were of high scores when evaluated with the PSSM model for T7-like promoters obtained in this work, the PHIRE predictions for the remaining six genomes have low PSSM scores (Supplementary Fig. S5A) and are likely to be false positives. For the same 46 genomes, PhageProm generates high-scoring predictions for 43 genomes (594 promoters in total), covering the 40 genomes for which PHIRE made high-scoring predictions. For the remaining three genomes where PhageProm did not identify any promoters (CR8, Percy, and ECBP5), the PSSM scores of the promoters identified by PHIRE were all also low.

When compared against PhagePromoter, a major advantage of PhageProm is that PhageProm but not PhagePromoter can discover promoters that diverge relatively far away from the “standard” T7 promoter motif. For instance, PhagePromoter did not predict any of the four experimentally verified promoters in the VSW-3 phage genome, which have sequence identities of 48% with the “standard” T7 promoter sequence [20]. In comparison, both PhageProm and PHIRE predicted five promoters in this phage genome, covering all the four experimentally confirmed promoters.

General features of T7-like promoters

Since T7-like promoters account for the majority of the discovered phage promoters, the sequence patterns of T7-like promoters warrant further analysis. For the 840 genomes containing predicted T7-like promoters, a total of 11218 promoters were predicted, averaging ∼13 promoters per genome. However, many predicted RNAP–promoter pairs share identical RNAP and/or promoter sequences. After removing redundancy, we obtained a total of 470 RNAPs and 1067 promoters of unique sequences. By matching the predicted promoters in a genome to the RNAP of the same genome, we obtained 4750 cognate RNAP–promoter pairs of unique sequences (Supplementary Dataset S1).

The 1067 T7-like promoters exhibit a clearly discernible consensus sequence pattern. As shown in Fig. 1B, the 21 bp promoter can be divided into three 5 bp subregions (regions I–III) and a 6 bp subregion (region IV). Among these, subregion III is the most conserved, while subregion II is the least conserved. Figure 1C shows the exact subsequence distributions in the different subregions. In subregion III, >50% of the promoters share the exact subsequence CACTA, while in the subregions I, II, and IV, the five most frequent subsequences account for 39.6, 47.7, and 25.1%, respectively, of the promoters. Supplementary Fig. S5B demonstrates a clear separation between the PSSM score distributions of predicted promoters and random genome fragments. Specifically, 99.3% of predicted promoters have scores above the crosspoint value, and 99.2% of random genome fragments have scores below the crosspoint value. This indicates high specificity and selectivity of the PSSM score as a metric for evaluating candidate promoters of the T7-like category.

Specificity determinants in T7-like RNAPs and promoters

We first tried to apply the mutual information (MI) metric to the protein–DNA joint MSA to infer specificity-determining amino acid residue and base positions. Although MI has been widely used to identify specificity-determining positions from MSAs of protein–DNA complexes [8, 9, 44], the metric does not differentiate between direct and indirect couplings. This could lead to too many false-positive predictions. Indeed, performing MI analysis on our data produced many positions on the RNAPs where residue types covaried with base types at various positions of the promoters. However, many of these positions were structurally distant from the promoter, as indicated by the experimental structure of the T7 RNAP–promoter complex [45] (Supplementary Fig. S6).

To overcome the problem associated with MI, we applied the DCA approach, which can filter out indirect coupling and thus lead to more accurate predictions than MI [46]. Previously, several DCA-based methods, including mfDCA, GREMLIN, and plmDCA [32, 47, 48], have been proposed for analyzing protein–protein interactions. Here we adjusted the plmDCA method to analyze protein–DNA interactions. We considered two possible models: bi-DCA and uni-DCA (Fig. 2A). For modeling the full-length T7-like RNAP–promoter MSA, the number of parameters of bi-DCA is ∼170 million, while the number of parameters of uni-DCA is ∼1.6 million. The former number was obviously too large relative to the limited number of sequences in our MSA. Therefore, we chose to use uni-DCA to identify strongly coupled residue–base position pairs in the full-length MSA of paired RNAPs and promoters.

Our initial results from uni-DCA analysis of the full-length RNAPs revealed a 28 residue region on the aligned RNAPs that displayed strong couplings with promoters (Fig. 2B). This region, named Region3, corresponds to the specificity loop of T7 RNAP, as revealed by the structure of the T7 RNAP–promoter complex [45]. However, a more detailed examination of the full-length uni-DCA results showed that the most strongly coupled amino acid residues and bases identified by the analysis did not match those involved in intermolecular hydrogen bonds in the complex structure (Supplementary Fig. S7A–C). This suggested that the full-length uni-DCA model could still lead to severe overfitting due to the limited amount of MSA data and the large number of model parameters. To mitigate overfitting, we explored two strategies. The first was to utilize a regularization term with a large weight when learning the full-length uni-DCA model (Supplementary Fig. S7D). The second was to consider only Region3 instead of the full-length RNAP in the uni-DCA analysis (Supplementary Fig. S7E). We found that both strategies could effectively mitigate overfitting, but the Region3-only model more accurately identified residue–base pairs that form intermolecular hydrogen bonds as the most strongly coupled pairs (Supplementary Fig. S7B). By restricting the analysis to Region3, we were also able to train a bi-DCA model without suffering from overfitting, with results in close agreement with those of the Region3-only uni-DCA model (Supplementary Fig. S7F). Consequently, we focused on the Region3-only bi-DCA results (Fig. 2C).

The results suggest that the residue–base pairs at 7 → 11 (numbers refer to the position on Region3 MSA → the position on the promoter), 9 → 7, 9 → 8, 17 → 8, 17 → 9, 18 → 10, and 19 → 10 dominantly determine the specificity of T7-like RNAP–promoter interactions. Figure 2D shows the frequency of each amino acid–base type combination at these position pairs. For example, for the position pair 9 → 7, the top three most frequent amino acid–base type combinations are 9D → 7C, 9N → 7G, and 9K → 7G. A special position pair is 7 → 11, where the amino acid–base type combination was highly conserved: the 7R → 11C combination was observed in 91% of cognate RNAP–promoter pairs.

As the specific recognition between T7-like RNAPs and promoters should rely on physical intermolecular interactions, we expect that the set of specificity-determining amino acid residues and bases derived from MSA-based analysis should generally agree with known structural data on the RNAP–promoter complexes. A previous structural study [45] revealed that the sequence-specific recognition of the T7 promoter by T7 RNAP is primarily achieved by the hydrogen-bonding interactions involving the side chains of four residues (R746, N748, R756, and Q758) in an antiparallel β-loop and bases in the major groove of the promoter. The specific residue–base pairs forming these hydrogen bonds are shown in Supplementary Fig. S7A. Indeed, the DCA of the MSA revealed that the residues at positions 748, 756, and 758 in this loop (corresponding to positions 9, 17, and 19 in Region3) could be critical for promoter specificity in other T7-like RNAPs as well. The convergence of the DCA and the structural data onto these positions gave us confidence in decoding specific RNAP–promoter interactions from the protein–DNA joint MSA based on the combinations of amino acid residue and base types at these positions.

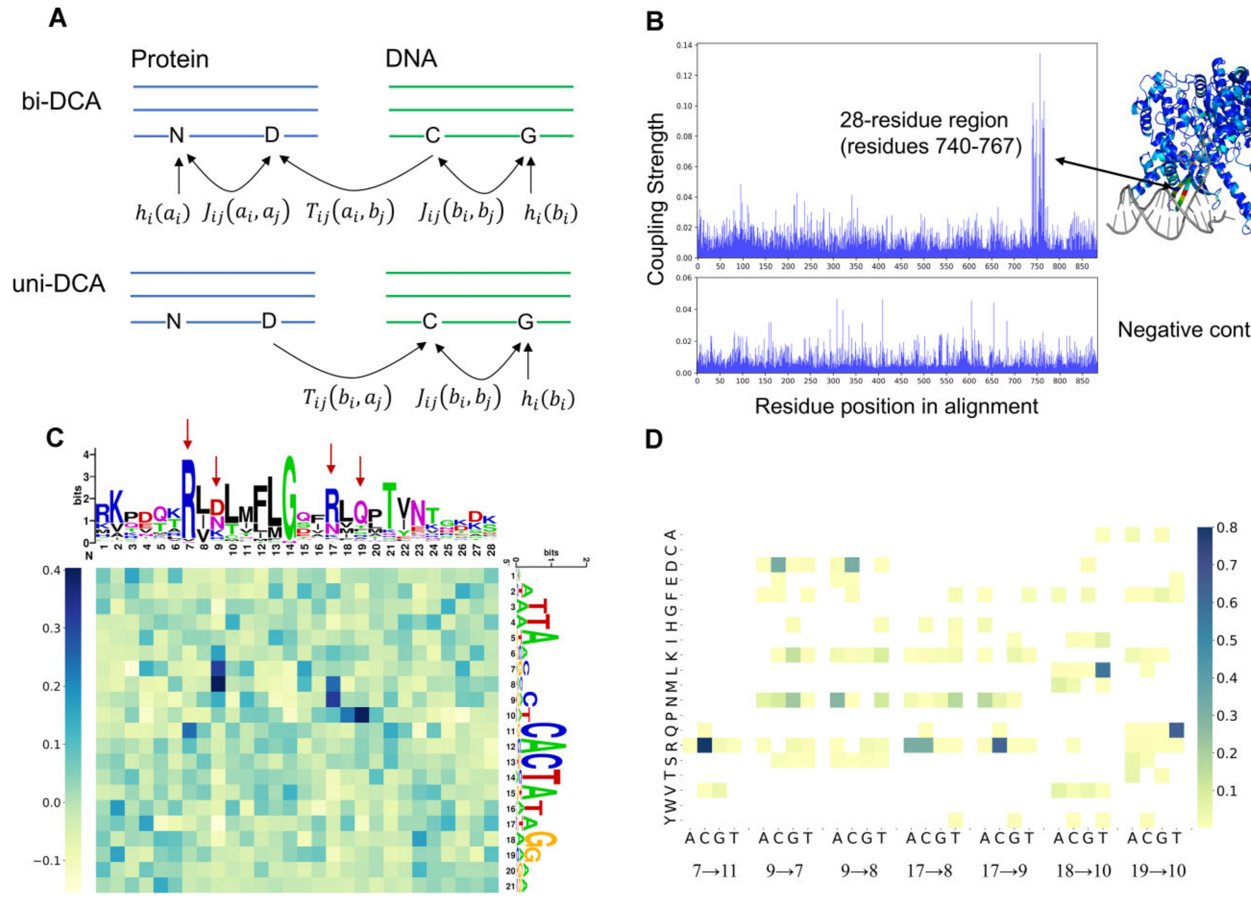


Figure 2. Direct coupling analysis of the joint MSA of T7-like RNAPs and cognate promoters. (**A**) The bi-DCA and uni-DCA models for modeling a protein–DNA joint MSA. (**B**) Left: the maximum coupling strength of each residue position with base positions obtained with uni-DCA. The upper and lower panels show the actual and the negative control results, respectively. Right: structure of T7 RNAP–promoter complex (pdbid: 1cez) with residues colored by a rainbow spectrum according to their maximum coupling strengths. The structure was drawn with PyMOL [54]. (**C**) The score matrix of coupling strength between each position in Region3 MSA and each position in promoter MSA inferred by Region3-only bi-DCA. Red arrows indicate the most strongly coupled amino acid residues. (**D**) Correlations between amino acid residue types and their coupled base types at the most strongly coupled positions from Region3-only bi-DCA.

Visualizing the connections between residue type and base type combinations in a network

In total, we observed 24 unique amino acid type combinations at positions 9, 17, and 19 of Region3 in the MSA of T7-like RNAPs, representing the naturally evolved diversity of these key residues for recognizing specific promoters. Correspondingly, we divided the T7-like RNAPs into 24 groups. Supplementary Fig. S8 shows the logos of the promoters recognized by RNAPs in each group.

Besides the sequence logos, the recognition relationships between the amino acid residue type combinations at the key positions on the protein side (positions 9, 17, and 19 of Region3 in the MSA) and the base type combinations at the key positions on the DNA side (positions 6–10 of the promoter MSA) were summarized and visualized as a network. In this SpSN as shown in Fig. 3A, all RNAPs sharing the same combination of amino acid residue types at the designated positions were assigned to the same node. The combinations of the base types in the cognate promoters of these RNAPs form a cluster of nodes connected to the corresponding amino acid type combination node. These nodes fall into nine larger clusters, each named after a representative RNAP having the amino acid combination of the central node. Some base type combination nodes are shared by similar but distinct amino acid type combinations which restrained the placement of the latter nodes in the network graph, causing the nodes of similar amino acid residue combinations (probably representing evolutionarily closer RNAPs) to be close to each other according to the shortest paths in the graph (Fig. 3B). The base type combination nodes bridging two protein nodes, e.g. the AACCT node bridging the T7 and T3 nodes or the AATAT node bridging the T7 and K1F nodes, may play pivotal roles in the evolutionary differentiation of promoter specificity of T7-like RNAPs.

We then experimentally tested if the network, which involves only a small number of positions extracted from the full-length RNAPs and promoters, could guide the engineering of a particular RNAP, such as the T7 RNAP. We experimentally quantified each connection in the network with the background sequences of T7 RNAP and its cognate promoter (i.e. only the amino acid residues and bases at the designated key positions were mutated to combinations as specified by the nodes) using an *in vivo* transcription assay. The results (Supplementary Dataset S2) showed that more than half of the connections (74 out of 137) were associated with active transcriptions (GFP fluorescence intensity > 1000 a.u.). With

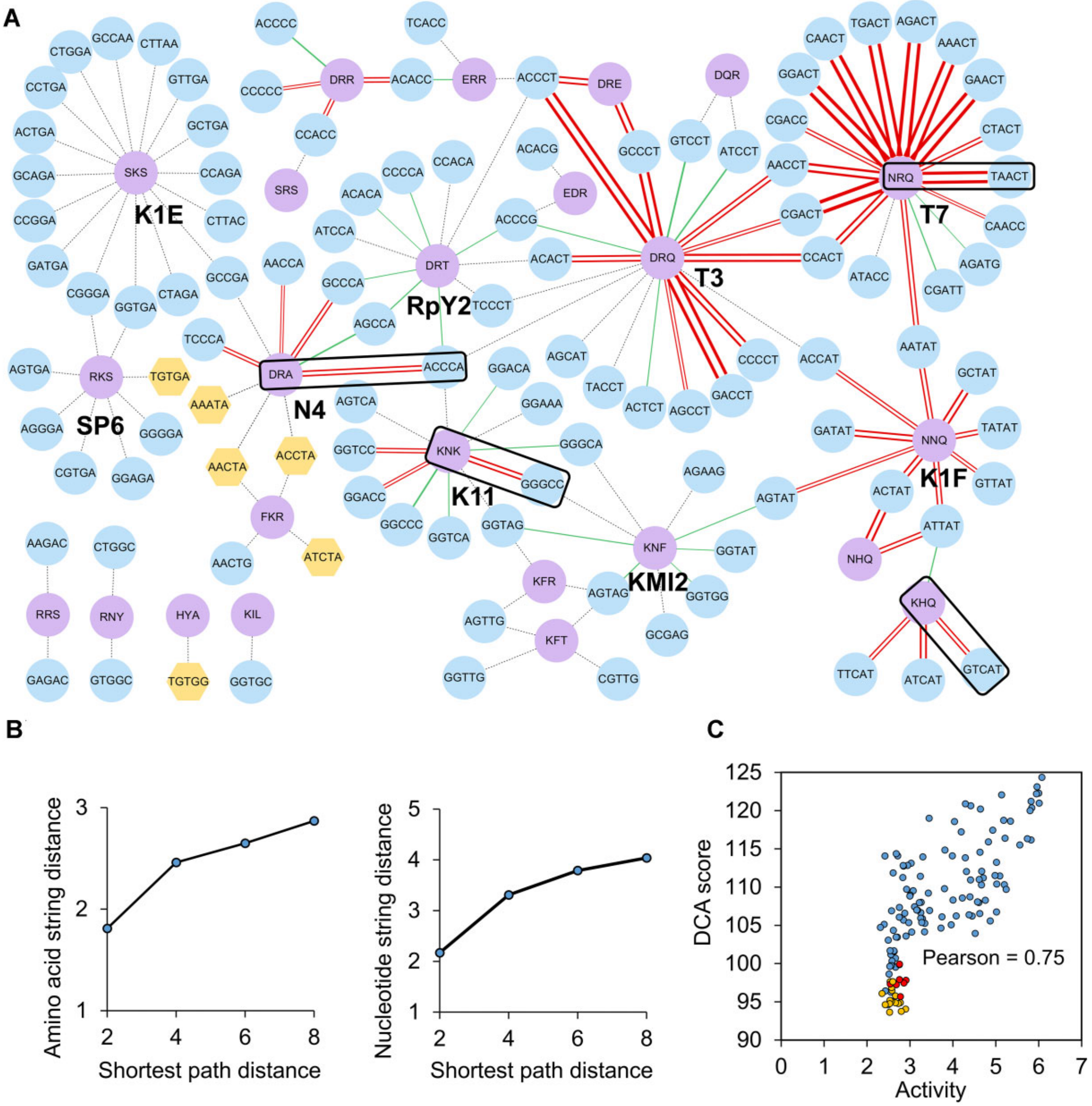


Figure 3. Summarized relationships between the amino acid type combinations at three key positions on T7-like RNAPs and the base type combinations at five key positions on promoters. (**A**) A network visualizing the relationships. Purple nodes represent amino acid residue type combinations. Blue and yellow nodes represent base type combinations. Edges represent observed pairings in natural T7-like RNAPs and cognate promoters. The double solid red edges indicate strong activities (GFP fluorescence intensity > 10 000 a.u.) observed in experimental tests of the corresponding pairs of T7 RNAP–promoter variants. The solid green edges indicate detectable but weak activities (GFP fluorescence intensity > 1000 a.u. but < 10 000 a.u.), with thicker lines indicating stronger activities. The yellow nodes are presumed to be promoters for host RNAPs because the GFP fluorescence intensity of the cells induced by IPTG (for expression of the T7 RNAP variants) was not significantly higher than that without IPTG induction (Supplementary Dataset S2). The four pairs used in the set of T7 RNAP variants with orthogonal promoter specificity are highlighted with rectangular boxes. (**B**) The relationship between the average shortest path distance and the string distance (i.e. the number of non-identical residues or bases) for two amino acid type combination nodes (left side) and two nucleotide type combination nodes (right side). (**C**) The correlation between the experimentally measured transcriptional activities and DCA scores of 137 pairs of T7 RNAP–promoter variants constructed according to the network. The activity values on the x-axis refer to the logarithm of GFP fluorescence intensity. The points colored in yellow and red correspond to combination pairs from the K1E and SP6 clusters, respectively.

the exception of the K1E and SP6 clusters, the other seven large clusters in the SpSN network have at least one pair of experimentally verified active RNAP–promoter variants. The K1E and SP6 RNAPs are the most evolutionarily distant ones from T7 RNAP. Therefore, changing only the three key residues in the T7 RNAP background sequence may not be sufficient to confer the specificity of K1E and SP6 RNAPs to T7 RNAP. We note that the transcriptional activity indirectly reflects protein–DNA binding strengths for the T7 RNAP–promoter variants, as previous studies have suggested that single or multi-amino acid substitutions on the specificity loop of T7 RNAP usually did not cause significant changes of solubility, expression and catalytic activity [3, 4]. Moreover, a previous study has shown that the transcriptional activi-

Table 1. A designed set of four orthogonal T7 RNAP–promoter variant pairs

	TAACT[a]	GTCAT	GGGCC	ACCCA
NRQ[a]	**6.41** (± 5.45)[b]	0 (± 1.30)[c]	3.51 (± 2.89)	2.63 (± 1.67)
KHQ	0 (± 1.82)	**5.37** (± 5.05)	1.65 (± 1.93)	2.68 (± 2.31)
KNK	2.51 (± 2.12)	1.98 (± 1.39)	**5.90** (± 5.33)	2.62 (± 2.29)
DRA	2.69 (± 2.44)	2.23 (± 1.79)	0 (± 1.25)	**5.29** (± 4.44)

[a] The amino acid and base type combinations. NRQ is the amino acid combination in the wild-type T7 RNAP.

[b] The activity values shown are the logarithm of the average of three independent cultures measured under 50 μM IPTG induction, after subtracting the activity value of the negative control. The values in parentheses indicate the logarithm of one standard deviation. The negative control was measured in the absence of the promoter. The highest activity values of each T7 RNAP variant on different promoters are highlighted in bold.

[c] The value '0' in the table means that the activity measured was lower than that of the negative control.

Table 2. Sets of orthogonal T7 RNAP–promoter variant pairs reported by various studies

Method	T7 RNAP variant	Cognate promoter[a]	Orthogonality[b]
Domain grafting [16]	632S	CGACT	8
	632S, 745K, 748D, 749M, 750I	ACCCT	
	632S, 754S, 756N, 761V	ACTAT	
	632S, 747I, 748D, 749C, 750V, 751I, 754T, 755H, 757M, 758A, 759L	ACCCA	
CPR [29]	Wild type	CGACT	34
	744K, 747V, 748H, 749I, 756E, 757M, 772R	CCGGT	
	725A, 744K, 747I, 748S, 749I, 756T, 758K, 772R, 775V	CCTGA	
	747I, 748D, 749C, 750V, 751I, 754T, 755R, 757M, 758A, 759L, 770N,772R, 775V	ACCCA	
DCA (this work)	Wild type	TAACT	245
	748K, 756H	GTCAT	
	748K, 756N, 758K	GGGCC	
	748D, 758A	ACCCA	

[a] Only the sequence of positions 6–10 of the promoter is shown, and the rest is the wild-type T7 promoter sequence.

[b] Orthogonality is defined as the smallest ratio among the activity values of the four cognate pairs divided by the maximum activity of their corresponding RNAP variants with non-cognate promoters.

ties of T7 RNAP and various promoter variants measured *in vivo* were closely correlated with their protein–DNA binding affinity [49].

We further scored the 137 experimentally tested T7 RNAP–promoter variant pairs using the bi-DCA model (Supplementary Methods) and compared the computed scores with their transcriptional activities (Fig. 3C). The Pearson correlation coefficient was 0.75. Notably, the DCA scores for all RNAP–promoter variant pairs in the K1E (yellow points) and SP6 (red points) clusters are significantly lower (Fig. 3C), indicating that the K1E-like and SP6-like RNAPs may deviate substantially from the other T7-homologous RNAP family members in their promoter specificity determination rules.

Overall, our experimental results verified that the combinations of amino acid residue types at only three key positions and the base types at only five key positions dominantly encode the interaction specificities between most T7-like RNAPs and their cognate promoters. Thus, it is possible to reprogram the promoter specificity of a wild-type RNAP by substituting only these amino acid residues/bases.

Designing orthogonal pairs of T7 RNAP–promoter variants

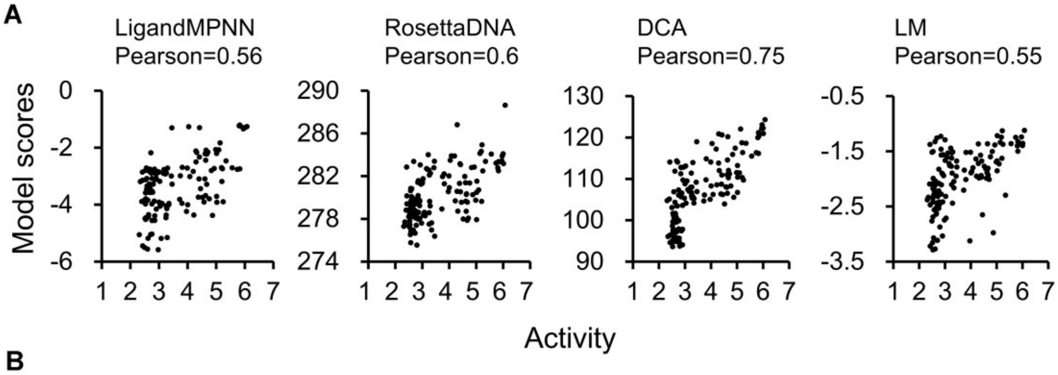
From the experiments described above, we identified 49 pairs of highly active T7 RNAP–promoter variants (indicated by the red double solid lines in Fig. 3A). From these pairs, it should be possible to select a set of RNAP–promoter variants with orthogonal specificity (i.e. of high activity between the cognate RNAP–promoter pairs and low activity for all noncognate pairs). To demonstrate this, we selected four pairs of highly active T7 RNAP–promoter variants, namely NRQ–TAACT, KHQ–GTCAT, DRA–ACCCA, and KNK–GGGCC (highlighted by the rectangular boxes in Fig. 3A). Within these sets, amino acid type combinations differ from each other by at least two residues, and the base type combinations differ by at least four bases. This design minimizes the likelihood of high activity in non-cognate pairs. The results of the *in vivo* transcription assays confirmed that these four pairs were highly orthogonal, with the transcription activity of cognate pairs > 200-fold higher than that of non-cognate pairs (Table 1; Supplementary Dataset S2). The cognate pairs NRQ-TAACT and KHQ-GTCAT exhibited the most stringent orthogonality, with almost undetectable activity observed in the non-cognate pairs (NRQ–GTCAT and KHQ–TAACT).

We note that previous studies have reported 4–6 pairs of orthogonal T7 RNAP—promoter variants constructed using domain grafting and compartmentalized partnered replication (CPR) methods, respectively [16, 29]. From the six pairs constructed in the CPR study, we selected the four with the most stringent orthogonality and compared them with the four pairs designed here. Results in Table 2 suggest that the pairs designed here show significantly stricter cross-pair orthogonality than previously reported pairs. We acknowledge that while the T7 RNAP–promoter variant pairs in Table 1 are stringently orthogonal, their absolute transcription activities are only ∼1/10 that of the wild-type T7 RNAP–promoter pair. However, as the activity of the wild-type T7 RNAP–promoter pair is often too high in intracellular applications, it may cause cellular toxicity without being turned down [16, 26, 50, 51]. Therefore, the lower activity of the reprogramed RNAP–promoter pairs should not be a significant limitation in applications.

Designing T7 RNAP–promoter variants with new combinations of interactions

Although we have shown that a large number of cognate T7 RNAP–promoter variant pairs predicted by DCA are indeed active, the amino acid residue type combinations and base type combinations in these pairs were all derived from natural sequences. Next, we examined if sequences with new type combinations at both the protein and the DNA sides could be computationally designed to achieve specific recognition.

We first tested the performance of two structure-based models for redesigning protein–DNA interactions, RosettaDNA [52] and LigandMPNN [53], on the T7 RNAP–promoter complex by examining how the scores produced by these methods correlate with the experimentally measured results of 137 RNAP–promoter variant pairs (Fig. 4A; Supplementary Methods). Both methods yielded reasonable Pearson correlation coefficients (0.6 for RosettaDNA and



B

9			17			19		
D	N	K	H	N	R	A	K/R	Q/E
6, 7, 8			8, 9			10		
AAC	AAA	ATC	CA	AC	AC	A	C	T
ACA	AAC	GGA	TA	CA	CC			
ACC	AAT	GGG		GC	CG			
AGC	ACC	GGT		TA	CT			
ATC	ACT	GTC		TC	TA			
CCA	AGA	TTC						
CCC	AGT							
CGA	ATT							
GAC	CAA							
GCC	CCA							
TCC	CGA							
	CTA							
	GAA							
	GAT							
	GCT							
	GGA							
	GTT							
	TAA							
	TAT							
	TGA							

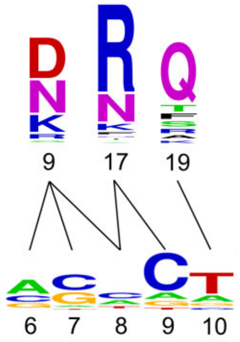


Figure 4. Designing T7 RNAP variants for recognizing new promoters. (**A**) Correlations between the scores of two structure-based models and two MSA-based models with the transcriptional activity data measured for 137 pairs of T7 RNAP–promoter variants. The activity values on the x-axis represent the logarithm of GFP fluorescence intensity. (**B**) The look-up table and mapped sequence logos illustrating the mapping from individual residue types to subgroups of base type combinations. The top two rows in the look-up table are the positions and residue types in Region3, while the bottom two rows are the mapped positions and base type combinations in the promoter.

0.56 for LigandMPNN). Notably, the RosettaDNA scores contained only contributions from the backbone-dependent residue type terms and the side chain hydrogen bond terms, rather than the total scores (see Supplementary Methods and Supplementary Table S3). However, these correlations were weaker than that computed with the DCA model (Pearson correlation coefficient 0.75). For comparisons, a language model trained on the RNAP–promoter sequence set (Supplementary Methods) produced a Pearson correlation coefficient of 0.55 (Supplementary Fig. S9). Interestingly, the Pearson correlation coefficient between the structure-based RosettaDNA scores and the MSA-based DCA scores was 0.77 (Table 3). Thus, in our setup, the evolutionarily selected protein–DNA pairs are also those that result in physically favorable interactions (i.e. strong protein–DNA hydrogen bonds) under the given structure.

Table 3. Pearson correlation coefficients between the scores of different models on 137 pairs of T7 RNAP–promoter variants

Method	RosettaDNA	DCA	LM
LigandMPNN	0.53	0.58	0.51
RosettaDNA	–	**0.77**	0.63
DCA	–	–	0.71

The highest value is in bold text.

To facilitate further designs, we simplified the DCA model into a look-up table based on the 49 experimentally determined RNAP–promoter variant pairs associated with high activity (Fig. 4B). In this table, each of the three key amino acid residues is coupled with only a subset of the five key bases. The allowed residue types are enumerated for each position, and each of them is mapped to a group of cognate base type combinations observed in the active RNAP–promoter variant pairs. Based on this table, new three-residue combinations can be generated from the allowed residue types, while the corresponding five-base combinations can be derived from the listed mappings from the residue type to the base type combinations. It is also possible to decode the residue type combination to recognize a given promoter sequence. Thus, the look-up table can be used as a guidance for generating new candidate cognate pairs.

We systematically considered all possible combinations of bases at the five key promoter positions, which comprised a total of $4^5 = 1024$ possibilities. After excluding combinations already covered by natural promoters, 917 possibilities remained. For 148 of them, corresponding residue type combinations at the three key RNAP sequence positions could be decoded according to the look-up table. This led to a total of 389 T7 RNAP–promoter variant pairs. In 152 pairs, the residue type combinations are new and all the five bases are different from those in the wild-type T7 promoter. Then, we applied RosettaDNA to evaluate the pairs and obtained 20 pairs with the highest scores. These 20 pairs involved 13 new promoter sequences and seven T7 RNAP variants with new residues type combinations at the three key positions, none of which had been observed in natural proteins. We experimentally determined the transcriptional activities of these 20 pairs. For comparison, we also determined the activities of the wild-type T7 RNAP on the 13 new promoter sequences. The results are presented in Table 4, and the raw fluorescence intensity data are provided in Supplementary Dataset S2. Most of the designed T7 RNAP variants showed higher activities on their corresponding target promoters than the wild-type T7 RNAP, with seven out of 20 pairs showing activities at least 10-fold higher than those of the wild type. These results confirmed the effectiveness of our approach of using the look-up table derived from the MSA to sample or propose designs and using a structure-based model such as RosettaDNA for engineering new specific interactions not yet sampled by natural evolution.

Table 4. The activities of designed T7 RNAP–promoter variants with new specific interactions

Target promoter	RNAP variant	RNAP variant activity[a]	T7 RNAP activity[a]	Selectivity[b]
AACGC	NRR	4.50 (± 3.82)	2.62 (± 1.73)	**74**
AACTC	DRK	3.02 (± 2.38)	2.17 (± 1.63)	7
AACTC	NRK	3.62 (± 2.87)	2.18 (± 1.63)	**27**
AACTC	NRR	3.66 (± 2.46)	2.18 (± 1.63)	**30**
ACCAC	DNK	3.18 (± 2.11)	2.63 (± 1.52)	4
ACCAC	DNR	2.44 (± 1.74)	2.63 (± 1.52)	1
ACCGC	DRK	3.79 (± 2.68)	2.74 (± 1.43)	**11**
ACCTC	DRK	3.62 (± 2.74)	2.27 (± 1.69)	**22**
GACAC	DNK	2.95 (± 1.85)	3.28 (± 2.19)	0
GACAC	DNR	2.48 (± 1.56)	3.28 (± 2.19)	0
GACGC	DRK	3.93 (± 2.34)	3.07 (± 2.57)	7
GACTC	DRK	2.17 (± 1.81)	2.24 (± 1.43)	1
GCCAC	DHK	4.58 (± 3.18)	2.87 (± 1.95)	**51**
GCCAC	DHR	3.49 (± 2.78)	2.87 (± 1.95)	4
GCCAC	DNK	3.32 (± 2.64)	2.87 (± 1.95)	3
GCCAC	DNR	2.69 (± 1.99)	2.87 (± 1.95)	1
GCCGC	DRK	3.72 (± 2.95)	2.87 (± 1.96)	7
GCCTC	DRK	3.49 (± 2.70)	2.42 (± 1.62)	**12**
TCCAC	DNK	2.98 (± 2.15)	2.74 (± 1.44)	2
TCCTC	DRK	3.14 (± 2.50)	2.45 (± 1.85)	5

[a] The activity values are the logarithm of the average of three independent cultures measured under 50 μM IPTG induction, after subtracting the activity value of negative control. The values in parentheses indicate the logarithm of one standard deviation.
[b] The selectivity is defined as the ratio of the activity of the RNAP variant on the target promoter to the activity of the wild-type T7 RNAP on the target promoter. The selectivity values >10 are highlighted in bold.

Discussion

We developed the PhageProm algorithm for the systematic identification of ssRNAPs and their recognized promoters within phage genomes. For the prediction of T7-like promoters, PhageProm demonstrates higher accuracy than PHIRE and greater diversity of predicted promoters compared with PhagePromoter. For the other three categories of ssRNAP promoters (phiKMV-like, KP34-like, and Syn5-like), it is difficult for PHIRE and PhagePromoter to make reliable predictions. For genomes without definite prediction of the promoters, we provided a small number of potentially promoter-containing subsequences for each genome, which could narrow the search range for future promoter characterization by experiments. The comprehensive predictions of T7-like promoters allowed us to train a DCA model to infer co-evolving pairs of protein residues and DNA bases from the RNAP–promoter joint MSA.

Our MSA provided systematic insights into the specificity of T7-like RNAP–promoter interactions. Over the past three decades, there has been a continuous effort to understand how promoter specificity is achieved by T7 RNAP [3, 4, 45]. These biochemical and structural studies have revealed the residues and bases at the key positions that determine the recognition specificity of T7 RNAP–promoters, as well as the interaction relationships between these key residues and bases. However, these studies have mainly focused on the T7 RNAP alone, not considering the large protein family. It is currently unclear which other family members follow the same recognition rules as T7 RNAP and which do not. Our analysis revealed that most T7-like RNAPs share a common structural basis for promoter recognition with T7 RNAP, while some others, such as K1E and SP6, did not follow the recognition rules derived from the DCA. The latter finding contrasts with the prior assumption that SP6 and T7 shared the same recognition rules [3, 4, 45].

Another important aspect of our analysis that goes beyond existing studies is that we can decode the combinations of residue/base types on the protein/DNA side and their connections based on a DCA model. Here the decoded information is visualized through a network, in which the shortest path distance between two nodes (RNAPs or promoters) is determined by their sequence similarity. These shortest path distances roughly correlate with evolutionary distances, which may support the gradual diverging of the promoter spectrum rather than abrupt switches of promoter specificity.

We demonstrated that our analysis can efficiently guide the engineering of promoter specificity for T7-like RNAP, especially for creating orthogonal RNAP–promoter pairs or designing RNAP–promoter pairs with new specific interactions. We showed that the orthogonal pairs designed here, based on DCA of the comprehensive family dataset, exhibited far lower off-diagonal activities than previous designs based on examining the sequence of a few individual family members.

We note that most of the transcriptional activities of the T7 RNAP–promoter variants characterized here are one order of magnitude or more lower than that of the wild-type T7 RNAP–promoter. This phenomenon has also been observed in previous studies [3, 4, 6, 16, 29]. Directed evolution experiments by previous researchers have yielded interesting results, showing that some mutations outside the specificity loop, such as E222K, H772R, and E775V, can generally improve the activities of these T7 RNAP variants without compromising their promoter recognition specificity [29, 31]. However, it is unclear whether these mutations promoted the binding of RNAPs to promoters, enhanced the expression or folding of RNAPs or increased the activity through some other unknown mechanisms. It is worthwhile addressing the questions of whether analyzing the full-length sequences of natural T7-like RNAPs can explain how these mutations function, and whether other approaches such as protein language models can be used to generate compensatory mutations to improve the activities of these T7 RNAP variants with altered promoter specificity.

Although the specific recognition between protein and DNA is eventually determined by intra- and inter-molecular physical interactions, existing structure-based models do not yet generally allow the accurate design of such recognition. Here we demonstrated that information derived from DCA of paired protein–DNA MSA can be combined with structure-based models such as RosettaDNA to efficiently design new protein–DNA interactions. The dataset and analysis results provided here may guide the promoter specificity engineering of T7 or T7-like RNAPs. Moreover, the approaches adopted here should be applicable to other protein–DNA recognition systems.

Acknowledgements

Author contributions Jinwei Zhu: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft. Haiyan Liu: conceptualization, formal analysis, methodology, supervision, writing—review & editing. Quan Chen: conceptualization, formal analysis, methodology, writing—review & editing. Chunbo Lou: conceptualization, formal analysis, methodology. Ziming Liu: conceptualization, methodology, validation.

Supplementary data

Supplementary Data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Key R&D Program of China [2022YFF1203100 to Q.C.]; the National Natural Science Foundation of China [92253302 to H.L. and 32371487 and 32171411 to Q.C.]; CAS Project for Young Scientists in Basic Research [YSBR-072 to Q.C.]; Anhui Provincial Natural Science Foundation [2308085J01 to Q.C.]; and Research Funds of Center for Advanced Interdisciplinary Science and Biomedicine of IHM [QYPY20230035 to Q.C.].

Data availability

All predicted promoters, subsequence groups, potential promoters with high PSSM scores, and T7-like RNAP–promoter pairs are available in Supplementary Dataset S1. The raw fluorescence intensity data of T7 RNAP–promoter variants are available in Supplementary Dataset S2. Python code for mining and analyzing T7-like RNAP–promoter sequence pairs is publicly available on Zenodo (https://doi.org/10.5281/zenodo.13982394).

References

1. Cermakian N, Ikeda TM, Miramontes P *et al.* On the evolution of the single-subunit RNA polymerases. *J Mol Evol* 1997;**45**:671–81. https://doi.org/10.1007/PL00006271
2. Krupp G. RNA synthesis: strategies for the use of bacteriophage RNA polymerases. *Gene* 1988;**72**:75–89. https://doi.org/10.1016/0378-1119(88)90129-1
3. Raskin CA, Diaz GA, McAllister WT. T7 RNA polymerase mutants with altered promoter specificities. *Proc Natl Acad Sci USA* 1993;**90**:3147–51. https://doi.org/10.1073/pnas.90.8.3147
4. Rong M, He B, McAllister WT *et al.* Promoter specificity determinants of T7 RNA polymerase. *Proc Natl Acad Sci USA* 1998;**95**:515–9. https://doi.org/10.1073/pnas.95.2.515
5. Dickinson BC, Leconte AM, Allen B *et al.* Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc Natl Acad Sci USA* 2013;**110**:9007–12. https://doi.org/10.1073/pnas.1220670110
6. Ellefson JW, Meyer AJ, Hughes RA *et al.* Directed evolution of genetic parts and circuits by compartmentalized partnered replication. *Nat Biotechnol* 2014;**32**:97–101. https://doi.org/10.1038/nbt.2714
7. Desai TA, Rodionov DA, Gelfand MS *et al.* Engineering transcription factors with novel DNA-binding specificity using comparative genomics. *Nucleic Acids Res* 2009;**37**:2493–503. https://doi.org/10.1093/nar/gkp079
8. Korostelev YD, Zharov IA, Mironov AA *et al.* Identification of position-specific correlations between DNA-binding domains and their binding sites. Application to the MerR family of transcription factors. *PLoS One* 2016;**11**:e0162681. https://doi.org/10.1371/journal.pone.0162681
9. Laforet M, McMurrough TA, Vu M *et al.* Modifying a covarying protein–DNA interaction changes substrate preference of a site-specific endonuclease. *Nucleic Acids Res* 2019;**47**:10830–41. https://doi.org/10.1093/nar/gkz866
10. Morgan RD, Luyten YA. Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. *Nucleic Acids Res* 2009;**37**:5222–33.https://doi.org/10.1093/nar/gkp535
11. Long P, Zhang L, Huang B *et al.* Integrating genome sequence and structural data for statistical learning to predict transcription factor binding sites. *Nucleic Acids Res* 2020;**48**:12604–17. https://doi.org/10.1093/nar/gkaa1134
12. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2024;**52**:D174–82. https://doi.org/10.1093/nar/gkad1059
13. Rodionov DA, Dubchak IL, Arkin AP *et al.* Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. *PLoS Comput Biol* 2005;**1**:e55. https://doi.org/10.1371/journal.pcbi.0010055
14. Roberts RJ, Vincze T, Posfai J *et al.* REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* 2023;**51**:D629–30. https://doi.org/10.1093/nar/gkac975
15. Durrant MG, Fanton A, Tycko J *et al.* Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol* 2023;**41**:488–99. https://doi.org/10.1038/s41587-022-01494-w
16. Temme K, Hill R, Segall-Shapiro TH *et al.* Modular control of multiple pathways using engineered orthogonal T7 polymerases. *Nucleic Acids Res* 2012;**40**:8773–81. https://doi.org/10.1093/nar/gks597

17. Zhao H, Zhang HM, Chen X *et al.* Novel T7-like expression systems used for Halomonas. *Metab Eng* 2017;**39**:128–40. https://doi.org/10.1016/j.ymben.2016.11.007
18. Cook R, Brown N, Redgwell T *et al.* INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. *Phage (New Rochelle)* 2021;**2**:214–23.
19. Lammens EM, Feyaerts N, Kerremans A *et al.* Assessing the orthogonality of phage-encoded RNA polymerases for tailored synthetic biology applications in Pseudomonas species. *Int J Mol Sci* 2023;**24**:7175. https://doi.org/10.3390/ijms24087175
20. Xia H, Yu B, Jiang Y *et al.* Psychrophilic phage VSW-3 RNA polymerase reduces both terminal and full-length dsRNA byproducts in in vitro transcription. *RNA Biol* 2022;**19**:1130–42. https://doi.org/10.1080/15476286.2022.2139113
21. Lavigne R, Sun WD, Volckaert G. PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. *Bioinformatics* 2004;**20**:629–35. https://doi.org/10.1093/bioinformatics/btg456
22. Sampaio M, Rocha M, Oliveira H *et al.* Predicting promoters in phage genomes using PhagePromoter. *Bioinformatics* 2019;**35**:5301–2. https://doi.org/10.1093/bioinformatics/btz580
23. Klucar L, Stano M, Hajduk M. phiSITE: database of gene regulation in bacteriophages. *Nucleic Acids Res* 2010;**38**:D366–70. https://doi.org/10.1093/nar/gkp911
24. Lu X, Wu H, Xia H *et al.* Klebsiella phage KP34 RNA polymerase and its use in RNA synthesis. *Front Microbiol* 2019;**10**:2487. https://doi.org/10.3389/fmicb.2019.02487
25. Putzeys L, Wicke L, Boon M *et al.* Refining the transcriptional landscapes for distinct clades of virulent phages infecting *Pseudomonas aeruginosa*. *Microlife* 2024;**5**:uqae002. https://doi.org/10.1093/femsml/uqae002
26. Segall-Shapiro TH, Meyer AJ, Ellington AD *et al.* A ‘resource allocator’ for transcription based on a highly fragmented T7 RNA polymerase. *Mol Syst Biol* 2014;**10**:742. https://doi.org/10.15252/msb.20145299
27. Shis DL, Bennett MR. 2013; Library of synthetic transcriptional AND gates built with split T7 RNA polymerase mutants. *Proc Natl Acad Sci USA* **110**:5028–33.
28. Pu J, Dewey JA, Hadji A *et al.* RNA polymerase tags to monitor multidimensional protein–protein interactions reveal pharmacological engagement of bcl-2 proteins. *J Am Chem Soc* 2017;**139**:11964–72. https://doi.org/10.1021/jacs.7b06152
29. Meyer AJ, Ellefson JW, Ellington AD. Directed evolution of a panel of orthogonal T7 RNA polymerase variants for in vivo or in vitro synthetic circuitry. *ACS Synth Biol* 2015;**4**:1070–6. https://doi.org/10.1021/sb500299c
30. Tracewell CA, Arnold FH. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. *Curr Opin Chem Biol* 2009;**13**:3–9. https://doi.org/10.1016/j.cbpa.2009.01.017
31. Carlson JC, Badran AH, Guggiana-Nilo DA *et al.* Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat Chem Biol* 2014;**10**:216–22. https://doi.org/10.1038/nchembio.1453
32. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J Comput Phys* 2014;**276**:341–56. https://doi.org/10.1016/j.jcp.2014.07.024
33. Yang H, Ma Y, Wang Y *et al.* Transcription regulation mechanisms of bacteriophages: recent advances and future prospects. *Bioengineered* 2014;**5**:300–4. https://doi.org/10.4161/bioe.32110
34. Davanloo P, Rosenberg AH, Dunn JJ *et al.* 1984; Cloning and expression of the gene for bacteriophage T7 RNA polymerase. *Proc Natl Acad Sci USA* **81**:2035–9.
35. Krieg PA, Melton D. In vitro RNA synthesis with SP6 RNA polymerase. *Methods Enzymol* 1987;**155:** 397–415. https://doi.org/10.1016/0076-6879(87)55027-3
36. Morris CE, Klement JF, McAllister WT. Cloning and expression of the bacteriophage T3 RNA polymerase gene. *Gene* 1986;**41**:193–200. https://doi.org/10.1016/0378-1119(86)90098-3
37. Zhu B, Tabor S, Raytcheva DA *et al.* The RNA polymerase of marine cyanophage Syn5. *J Biol Chem* 2013;**288**:3545–52. https://doi.org/10.1074/jbc.M112.442350
38. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinf* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421
39. Ester M, Kriegel H-P, Sander J *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD-96 Proceedings*. 1996, 226–31.
40. Studier FW. Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *J Mol Biol* 1991;**219**:37–44. https://doi.org/10.1016/0022-2836(91)90855-Z
41. Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* 1986;**189**:113–30. https://doi.org/10.1016/0022-2836(86)90385-2
42. Sun Y, Xu J, Zhou H *et al.* Recombinant protein expression chassis library of *Vibrio natriegens* by fine-tuning the expression of T7 RNA polymerase. *ACS Synth Biol* 2023;**12**:555–64. https://doi.org/10.1021/acssynbio.2c00562
43. Tan SI, Ng IS. New insight into plasmid-driven T7 RNA polymerase in *Escherichia coli* and use as a genetic amplifier for a biosensor. *ACS Synth Biol* 2020;**9**:613–22. https://doi.org/10.1021/acssynbio.9b00466
44. Yang S, Yalamanchili HK, Li X *et al.* Correlated evolution of transcription factors and their binding sites. *Bioinformatics* 2011;**27**:2972–8. https://doi.org/10.1093/bioinformatics/btr503
45. Cheetham GM, Jeruzalmi D, Steitz TA. Structural basis for initiation of transcription from an RNA polymerase–promoter complex. *Nature* 1999;**399**:80–3. https://doi.org/10.1038/19999
46. Mao W, Kaya C, Dutta A *et al.* Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. *Bioinformatics* 2015;**31**:1929–37. https://doi.org/10.1093/bioinformatics/btv103
47. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 2013;**110**:15674–9. https://doi.org/10.1073/pnas.1314045110
48. Morcos F, Pagnani A, Lunt B *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 2011;**108**:E1293–1301. https://doi.org/10.1073/pnas.1111471108
49. Zhang X, Yao Z, Duan Y *et al.* Investigation of specific interactions between T7 promoter and T7 RNA polymerase by force spectroscopy using atomic force microscope. *Biochem J* 2018;**475**:319–28. https://doi.org/10.1042/BCJ20170616
50. Kar S, Ellington AD. Construction of synthetic T7 RNA polymerase expression systems. *Methods* 2018;**143**:110–20. https://doi.org/10.1016/j.ymeth.2018.02.022
51. Kushwaha M, Salis HM. A portable expression resource for engineering cross-species genetic circuits and pathways. *Nat Commun* 2015;**6**:7832. https://doi.org/10.1038/ncomms8832
52. Thyme S, Baker D. Redesigning the specificity of protein–DNA interactions with Rosetta. *Methods Mol Biol* 2014;**1123**: 265–82. https://doi.org/10.1007/978-1-62703-968-0_17
53. Dauparas J, Lee GR, Pecoraro R *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. bioRxiv, https://doi.org/10.1101/2023.12.22.573103, 23 December 2023, preprint: not peer reviewed.
54. Schrodinger L. The PyMOL molecular graphics system. Version 2015;**1**:8.

Received: November 1, 2024. **Revised:** February 2, 2025. **Editorial Decision:** February 10, 2025. **Accepted:** February 14, 2025