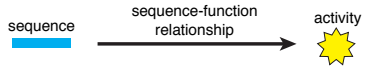
Quantitative sequence-function relationships

Justin B. Kinney

QB course
19 October 2018

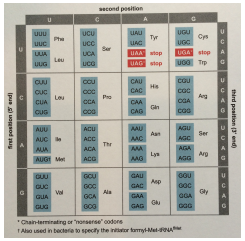
Sequence-function relationships are a fundamental phenomena of biology



sequence = DNA, RNA, protein, etc.

activity = anything the sequence does that is biologically relevant

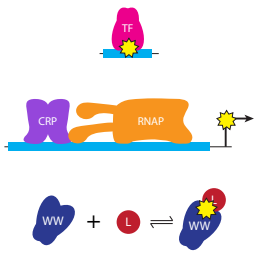
The genetic code is the best known sequence-function relationship



sequence	activity
mRNA ORF	protein
↓	↓
codon	amino acid or stop
(64 possibilities)	(21 possibilities)

The genetic code was solved by exhaustive enumeration

There are a wide variety of sequence-function relationships



sequence	activity
DNA binding site	TF binding energy
promoter DNA sequence	transcriptional activity
protein sequence	ligand binding affinity

Most sequence-function relationships are <u>quantitative!</u>

How do we represent quantitative sequence-function relationships?

The simplest way to represent a sequence-function relationship is enumerate the activities of all sequences

$4^{7.5}$ = 32,768 sequences

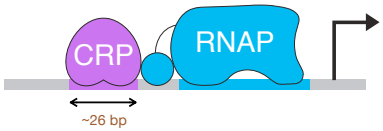
5'ss (9-mer)	PSI
AAAGCAAAA	37.3%
AAAGCAAAC	68.8%
AAAGCAAAG	75.6%
AAAGCAAAU	44.3%
AAAGCAACA	54.8%
AAAGCAACC	59.1%
AAAGCAACG	17.2%
AAAGCAACU	69.3%
AAAGCAAGA	49.6%
…	…

transcription + splicing

inclusion vs. exclusion

Exon inclusion for each 5'ss is quantified by <u>percent spliced in</u> (PSI)

Enumeration is possible only for very short sequences



4^{26} = 4.5 x10^{15} sequences

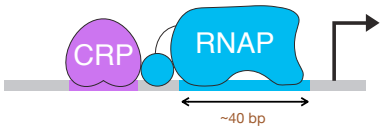
CRP site (26-mer)	kcal/mol
AAAAAAAAAAAAAAAAAAAAAAAAAA	0.26217
AAAAAAAAAAAAAAAAAAAAAAAAAC	0.22347
AAAAAAAAAAAAAAAAAAAAAAAAAG	0.18842
AAAAAAAAAAAAAAAAAAAAAAAAAT	0.23079
AAAAAAAAAAAAAAAAAAAAAAAACA	0.27212
AAAAAAAAAAAAAAAAAAAAAAAACC	0.20797
AAAAAAAAAAAAAAAAAAAAAAAACG	0.20228
AAAAAAAAAAAAAAAAAAAAAAAACT	0.2129
AAAAAAAAAAAAAAAAAAAAAAAAGA	0.24779
…	…

At 1 byte per measurement would require 1 petabyte of storage

Just sequencing the assayed DNA would require 10^{15} reads

A NovaSeq run provides ~1B = 10^{9} reads. Just sequencing this much DNA would require a million of these runs.

Enumeration is possible only for very short sequences



4^{40} = 10^{24} sequences

RNAP site (40-mer)	kcal/mol
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	0.26217
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC	0.22347
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAG	0.18842
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAT	0.23079
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACACA	0.27212
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACACC	0.20797
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACACG	0.20228
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACACT	0.2129
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAGA	0.24779
…	…

At 1 byte per measurement would require 1 million exabytes of storage

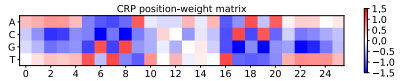
Digital storage capacity on Earth circa 2015 was ~300 exabytes.

The 10^{24} DNA molecules that would have to be measured would weight ~40 kg.

Position Weight Matrix (PWM): The simplest model for a quantitative sequence-function relationship

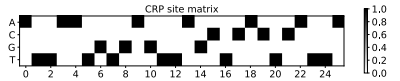
PWMs are used to assign a “score” to binding sites

w_{bi}



w_{bi} = weight contributed by base b at position i

s_{bi}



$$s_{bi} = \begin{cases} 1 & \text{if base b occurs at position i} \\ 0 & \text{otherwise} \end{cases}$$

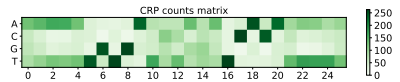
score assigned to a sequence: $W(s) = \sum_i \sum_b s_{bi} w_{bi}$

Higher-scoring sites bind more strongly.

Computing PWMs: counts matrix

list of aligned binding sites

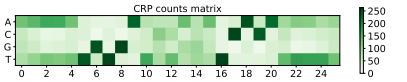
c_{bi}



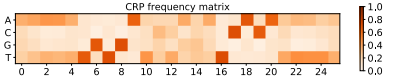
c_{bi} = number of sequences with base b at position i

Computing PWMs: frequency matrix

c_{bi}



f_{bi}



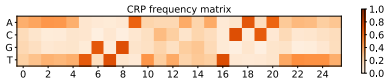
$$f_{bi} = \frac{c_{bi} + \lambda}{N + 4\lambda}$$

N = total number of aligned sequences

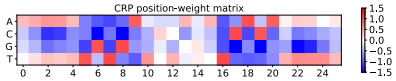
λ = pseudo count (typically 1 or 0.5)

Computing PWMs: the weight matrix

f_{bi}



w_{bi}



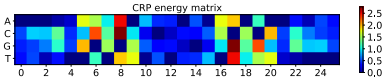
$$w_{bi} = \log \frac{f_{bi}}{p_b}$$

$p_b =$ background probability of base b

A PWM describes a generative model of binding sites

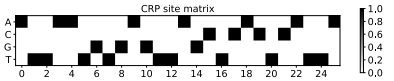
Energy matrix

g_{bi}



$g_{bi} =$ energy contribution of base b at position i

s_{bi}



$$s_{bi} = \begin{cases} 1 & \text{if base b occurs at position i} \\ 0 & \text{otherwise} \end{cases}$$

Gibbs free energy assigned to a sequence: $G(s) = \sum_i \sum_b s_{bi} g_{bi}$

An energy matrix describes protein-DNA binding energy

Relationship between PWMs and energy matrices

Stormo & Fields, 1998

$$g_{bi} = -k_B T \; w_{bi}$$

Berg & von Hippel, 1987

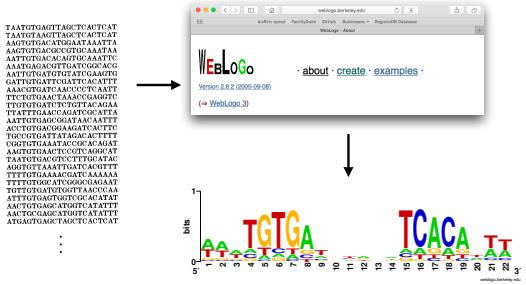
$$g_{bi} = -\alpha \; w_{bi}$$

$\alpha =$ unknown positive constant with units of energy

typically, $\alpha \sim 1 \; k_B T$

Berg & von Hippel, 1987, JBM
Stormo & Fields, 1998, TIBS

Sequence Logos



column height: information (in bits) = $I_i = \sum_b f_{bi} \log_2 \frac{f_{bi}}{p_b}$

Schneider & Stevens, 1990, NAR
https://weblogo.berkeley.edu/logo.cgi

Predicting binding sites within genomic sequences

Evaluate PWM score on all sites of the correct length

Call sites whose scores lie above some threshold

FIMO
Find Individual Motif Occurrences

http://meme-suite.org/doc/fimo.html

Motif finding: MEME

MEME
Multiple Em for Motif Elicitation
Version 5.0.2

input:
1. set of long sequences enriched in binding sites
2. model for random DNA

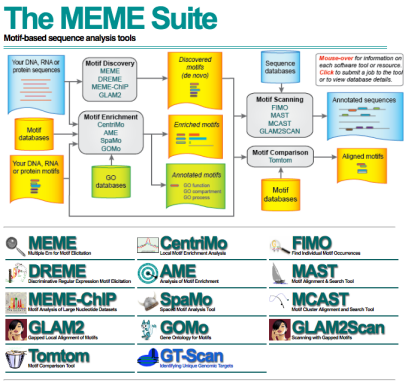
output:
1. PWM
2. location of putative sites

infer PWM ⟷ identify candidate sites within sequences

Expectation
Maximization (EM)
Algorithm

http://meme-suite.org/tools/meme

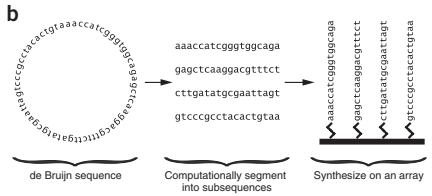
There is a lot of existing software for finding/using motifs



http://meme-suite.org/index.html

Technologies for measuring quantitative sequence-function relationships

Protein-binding microarrays

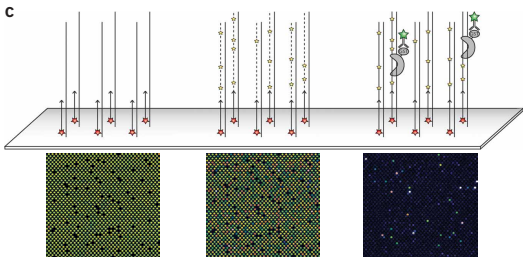


A de Bruijn sequence was designed with all 10-mers occurring exactly once

44K segments (60 bp each) covering this de Bruijn sequence were designed and printed as ssDNA on a microarray

Berger, et al., 2006, Nat. Biotechnol.

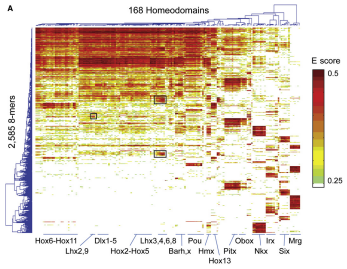
Protein-binding microarrays



ssDNA was made double-stranded. TF binding domains were then bound to the PBM and visualized with fluorescently labeled antibodies

Berger, et al., 2006, Nat. Biotechnol.

Protein-binding microarrays

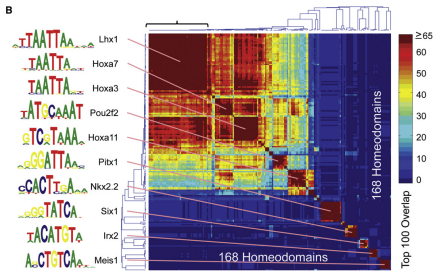


168 Homeodomains were assayed

An enrichment score was assigned to each 8-mer (which occurs on 16 different times)

Berger, et al., Cell, 2008

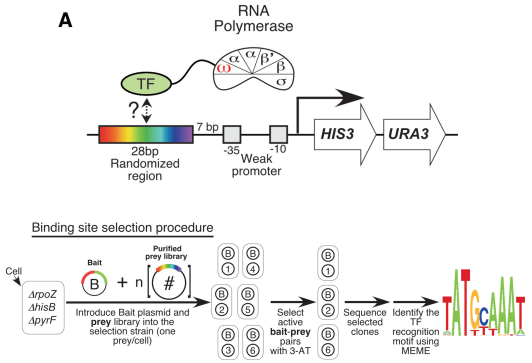
Protein-binding microarrays



Motifs were inferred from these enrichment scores, providing a characterization of most mouse homeodomain TFs

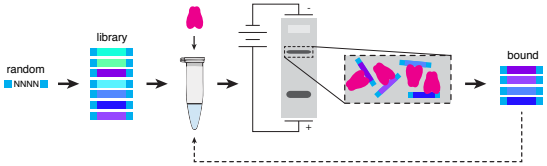
Berger, et al., Cell, 2008

E. coli one-hybrid



Noyes et al., 2008, NAR

SELEX-Seq

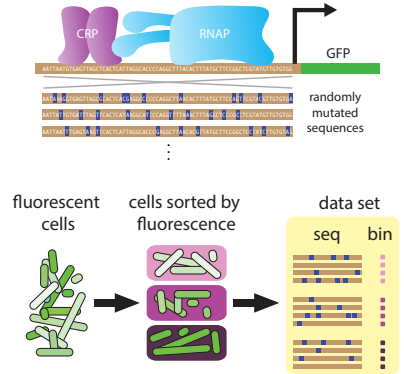


Library of random DNA is selected in vitro for TF-DNA binding

Comparison of bound DNA to input library yields TF motifs

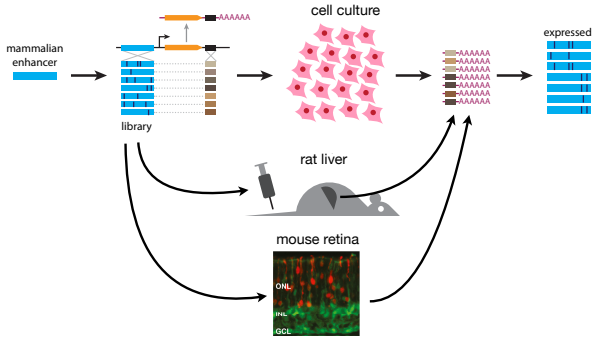
Slattery et al., 2011, Cell
(also: Zykovitch et al., 2009, NAR; Zhao et al., 2009; PLoS Comp. Biol., Jolma et al., 2009, Genome Res.; Wong et al., Genome Biol., 2011)

Massively parallel reporter assays: Sort-Seq



Kinney et al., 2010, PNAS

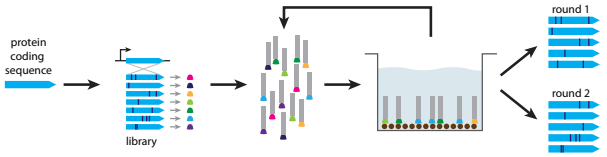
Massively parallel reporter assays: RNA-Seq



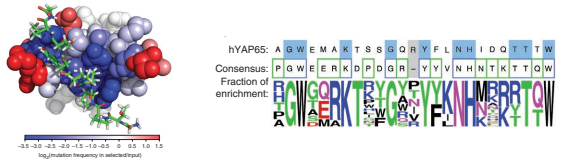
Melnikov et al., 2012, Nat. Biotechnol.

Patwardhan et al., 2012, Nat. Biotechnol.
Kwasnieski et al., 2012, PNAS

Deep mutational scanning

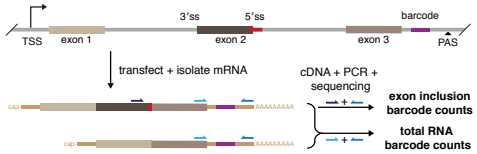


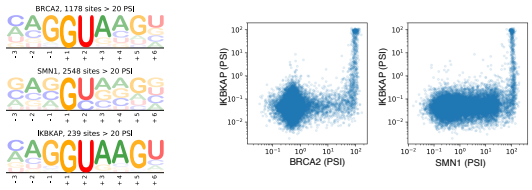
WW domain + peptide ligand



Fowler et al., 2010, Nat. Methods.

Massively Parallel Splicing Assay





Wong et al., 2018, Mol. Cell

References

Berg O, Hippel von P. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 1987;193(4):723–50.

Berger M, Philippakis A, Qureshi A, He F, Estep P, Bulyk M. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006 Nov 1;24(11):1429–35.

Berger M, Badis G, Gehrke A, Talukder S, Philippakis A, Pena-Castillo L, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 2008 Jun 27;133(7):1266–76.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010 Sep 1;7(9):741–6.

Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci USA. 2010 May 18;107(20):9158–63.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci USA. 2012 Nov 20;109(47):19498–503.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012 Feb 26;30(3):271–7.

Noyes M, Meng X, Wakabayashi A, Sinha S, Brodsky M, Wolfe S. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucl Acids Res. 2008 May 1;36(8):2547–60.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012 Feb 26;30(3):265–70.

Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucl Acids Res. 1990 Oct 25;18(20):6097–100.

Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, et al. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell. 2011 Dec 9;147(6):1270–82.

Stormo G, Fields D. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci. 1998;23(3):109–13.

Wong MS, Kinney JB, Krainer AR. Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. Mol Cell. 2018 Aug 16;71(6):1012–3.