Prospects for a Protein-DNA "Recognition Code"

Gary Stormo
ITP, UCSB
February 26, 2003

What is the value of a Protein-DNA recognition code?

Given a protein sequence, predict its binding sites and regulated genes in the genome.

Given a DNA binding site, predict which of the genome's proteins bind to it.

Design a protein to bind a specific DNA site, or DNA site to bind a specific protein.

Proc. Nat. Acad. Sci. USA
Vol. 73, No. 3, pp. 804–808, March 1976
Biophysics

Sequence-specific recognition of double helical nucleic acids by proteins

(base pairs/hydrogen bonding/recognition fidelity/ion binding)

NADRIAN C. SEEMAN, JOHN M. ROSENBERG*, AND ALEXANDER RICH

Department of Biology, Massachusetts Institute of Technology, Cambridge, Mass. 02139

Contributed by Alexander Rich, December 10, 1975

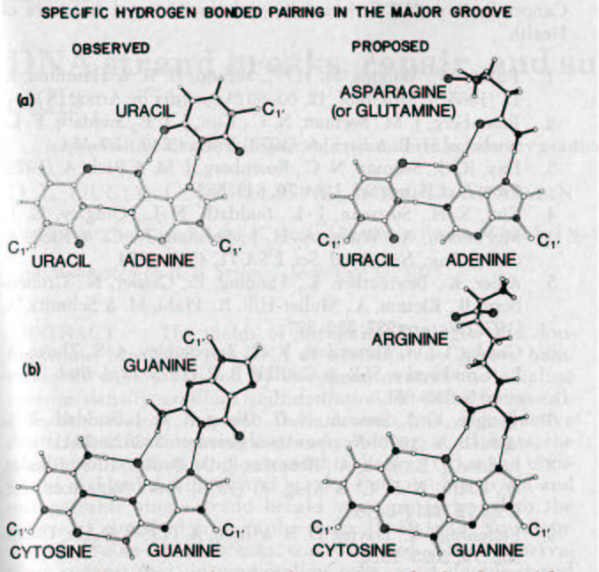


FIG. 2. Interactions between base pairs and other bases (observed) or amino-acid side chains (proposed). (a) The uracil binding to the U·A pair is seen in polynucleotides (11) as well as in single crystals of adenine and uracil derivatives (18). The conformation of the asparagine is taken from a neutron diffraction study (19). (b) The guanine binding to the C·G pair is seen in yeast phenylalanine tRNA (12, 13). The arginine conformation is taken from a neutron crystallographic analysis (20). Oxygen atoms have diagonal shading while nitrogen atoms are stippled.

Protein–DNA interaction

No code for recognition

Brian W. Matthews

THIS has been something of a banner year for repressor–operator complexes. On page 321 of this issue[1], Sigler and colleagues describe the structure of *trp* repressor complexed with its synthetic operator. Other complexes involving the DNA-binding domains of λ-repressor[2] and of phage 434 repressor[3] as well as 434 Cro protein[4] have also now been determined. The structure at near-atomic resolution of another complex of 434 repressor was described last year[5]. It is an opportune time to take stock of the field and to consider implications for the future.

that the binding geometry of the recognition helix on the DNA is similar for many repressor proteins [1-7]. The complexes now available show that the geometry of binding is variable. Even in the case of 434 Cro[4] and 434 repressor headpiece[3,5], two proteins with 50 per cent amino-acid sequence identity and very similar three-dimensional structures, and which are bound to the same operator, the respective helix-turn-helix units interact with the DNA in similar, but distinctly different, ways[4].

Specificity. The early studies of Cro, CAP the protein and the phosphate backbone contribute to the overall affinity of binding; sequence-specific contacts with the exposed parts of the operator base pairs allow the repressor to discriminate between its operator site and other DNA sequences.

Direct or indirect readout? The complexes involving λ-repressor headpiece, 434 Cro and 434 repressor headpiece all display multiple contacts both to the DNA backbone and to the parts of the base pairs that are exposed within the grooves of the DNA[2-5]. Although deviations from uniform B-form are observed in the conformations of the DNA, the recognition of the specific operator sequences appears to occur primarily by direct readout. In other words, the DNA retains essentially B-type conformation and the protein directly reads the sequence infor-

doi:10.1006/jmbi.2000.3918 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **301**, 597–624

JMB

Geometric Analysis and Comparison of Protein-DNA Interfaces: Why is there no Simple Code for Recognition?

Carl O. Pabo* and Lena Nekludova

Howard Hughes Medical Institute, Department of Biology 68-580, Massachusetts Institute of Technology Cambridge, MA 02139, USA

**Corresponding author*

Structural studies of protein-DNA complexes have shown that there are many distinct families of DNA-binding proteins, and have shown that there is no simple "code" describing side-chain/base interactions. However, systematic analysis and comparison of protein-DNA complexes has been complicated by the diversity of observed contacts, the sheer number of complexes currently available and the absence of any consistent method of comparison that retains detailed structural information about the protein-DNA interface. To address these problems, we have developed geometric methods for characterizing the local structural environment in which particular side-chain/base interactions are observed. In particular, we develop methods for analyzing and comparing spatial relationships at the protein-DNA interface. Our method involves attaching local coordinate systems to the DNA bases and to the C$^{\alpha}$ atoms of the peptide backbone (these are relatively rigid structural units). We use these tools to consider how the position and orientation of the polypeptide backbone (with respect to the DNA) helps to determine what contacts are possible at any given position in a protein-DNA complex. Here, we focus on base contacts that are made in the major groove, and we use spatial relationships in analyzing: (i) the observed patterns of side-chain/base interactions; (ii) observed helix docking orientations; (iii) family/subfamily relationships among DNA-binding proteins; and (iv) broader questions about evolution, altered specificity mutants and the limits for the design of new DNA-binding proteins. Our analysis, which highlights differences in spatial relationships in different complexes and at different positions in a complex, helps explain why there is no simple, general code for protein-DNA recognition.

Keywords: protein-DNA complexes; recognition code; recognition helix; major groove; α helix

Kinds of Codes

Deterministic (Simple, one-to-one)

Genetic Code: CCA _ Pro

Probabilistic (Degenerate, many-to-one)

Genetic Code (reversed)

Pro _ {CCA, CCC, CCG, CCT}

Probabilistic (many-to-many)

{cl, cro} _ {OR1, OR2, OR3,…}

Existing models...

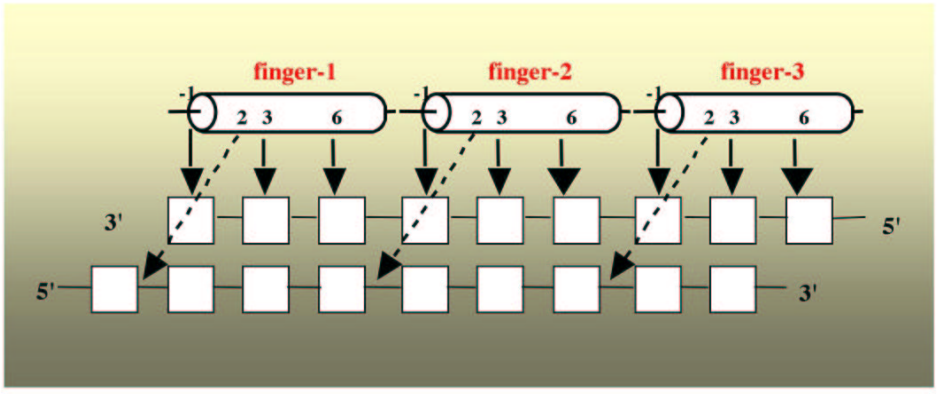
A. Qualitative models

- Desjarlais and Berg (1992)
- Pabo *et al.* (1997)
- Choo and Klug (1997)

B. Quantitative models

- Suzuki *et al.* (1995)
- Margalit *et al.* (1998)

EGR family of transcription factors



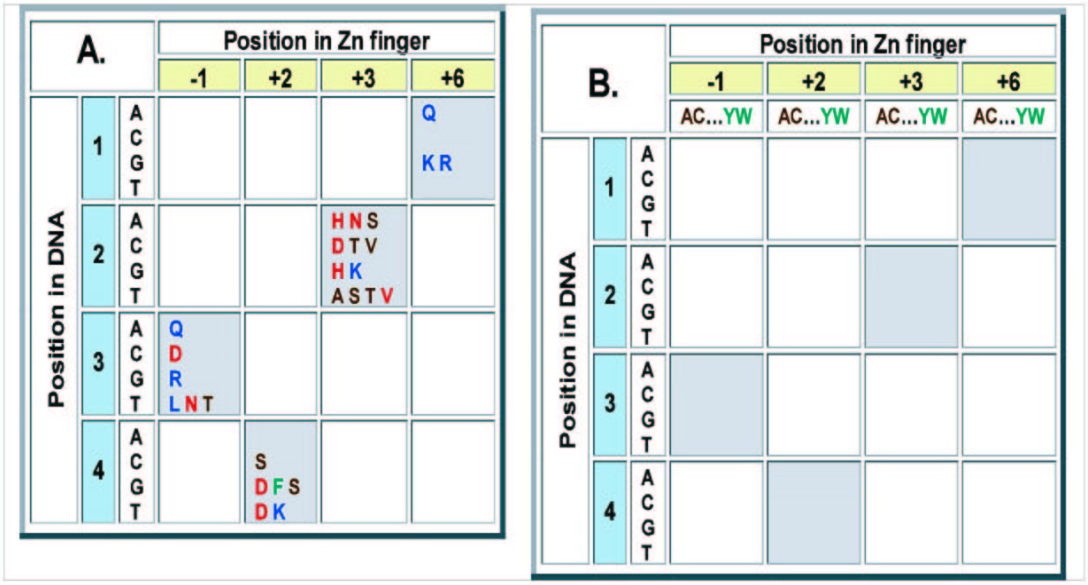
Zn-finger (Cys_2His_2):

- 3 fingers, binding in a modular fashion
- target site: 4 bases long (for each finger)
- one base overlap in the target of each finger

Qualitative models...

- Usually in the form of a simple "binary" table.
- Difficult to expand to other than "one-to-one" model.
- By nature, unsuitable for quantitative predictions.

	Position in triplet		
	5'	Middle	3'
A	Gln 6	Asn 3 Ser 3 His 3	Gln -1
C	Ser 2	Asp 3 Thr 3 Val 3	Asp -1
G	Arg 6 Lys 6 Asp 2 Ser 2 Phe 2	His 3 Lys 3	Arg -1
T	Lys 2 Asp 2	Thr 3 Ala 3 Ser 3 Val 3	Leu -1 Thr -1 Asn -1

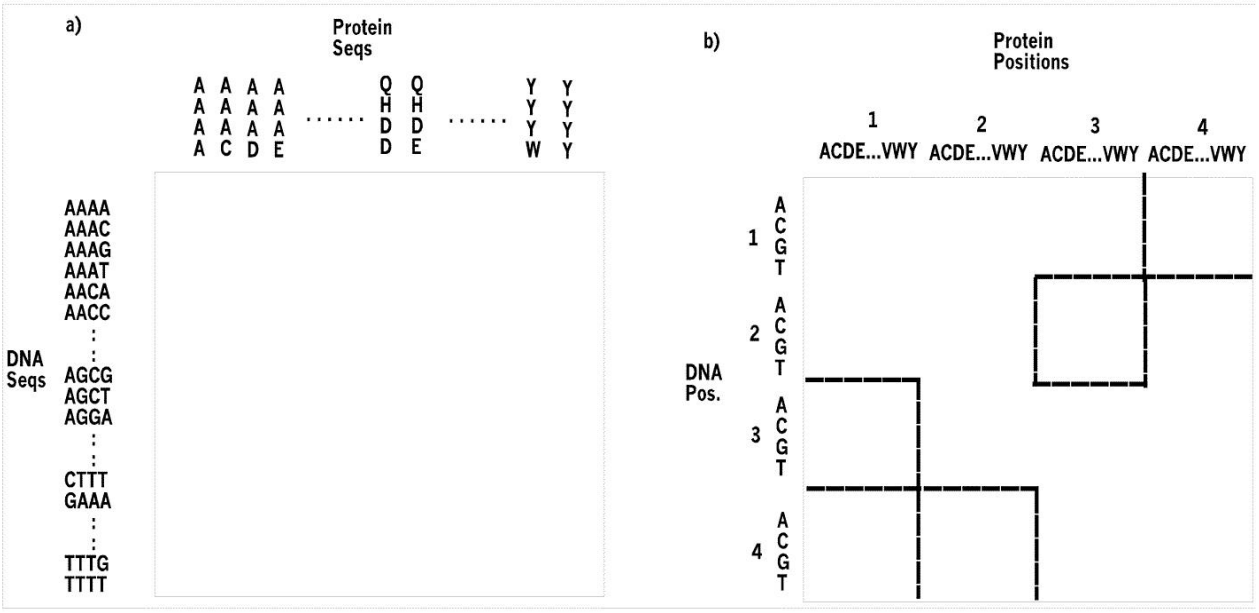


a)

AAAA	2.63
AAAC	2.63
AAAG	2.63
AAAT	0.46
AACA	1.79
AACC	1.79
AACG	1.79
AACT	-0.38
:	:
:	:
:	:
AGCG	-0.25
AGCT	-2.42
AGGA	0.59
:	:
:	:
:	:
CTTT	-0.37
GAAA	4.01
:	:
:	:
:	:
TTTG	2.63
TTTT	0.46

b)

	1	2	3	4
A	-0.55	+1.38	+0.42	+1.38
C	0	+0.55	-0.42	+1.38
G	+0.83	-0.66	+0.42	+1.38
T	+0.83	+0.42	0	-0.79



4442–4451 *Nucleic Acids Research, 2002, Vol. 30 No. 20*

Additivity in protein–DNA interactions: how good an approximation is it?

Panayiotis V. Benos[1,2], Martha L. Bulyk[3] and Gary D. Stormo[1,*]

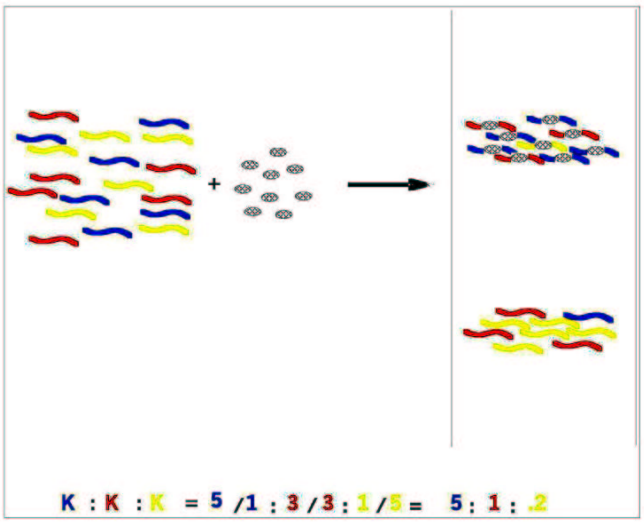
[1]Department of Genetics, Campus Box 8232, Washington University, School of Medicine, St Louis, MO 63110, USA, [2]Department of Human Genetics and Center for Computational Biology and Bioinformatics and Cancer Institute, University of Pittsburgh, PA 15261, USA and [3]Division of Genetics, Department of Medicine and Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, and Harvard-MIT Division of Health Sciences and Technology, Boston, MA 02115, USA

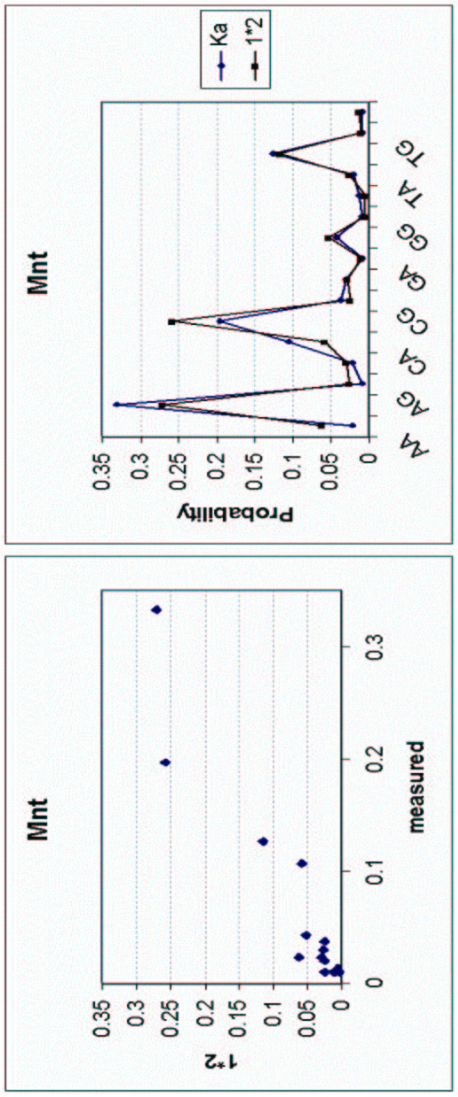
Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors

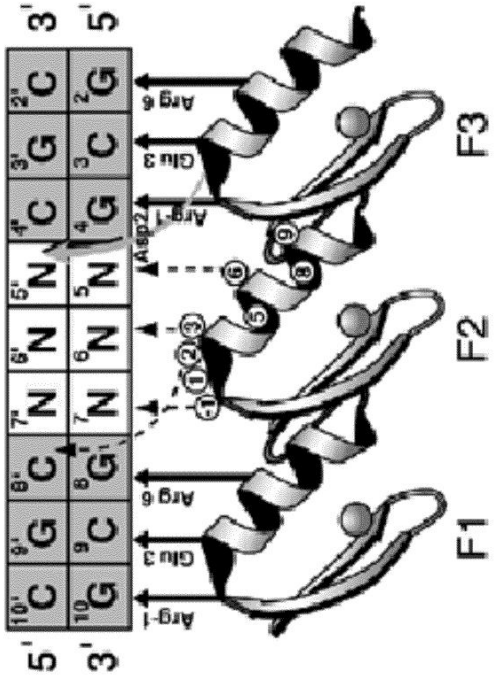
Martha L. Bulyk , Philip L. F. Johnson and George M. Church
Nucleic Acids Research, 2002, Vol. 30, 1255–1261

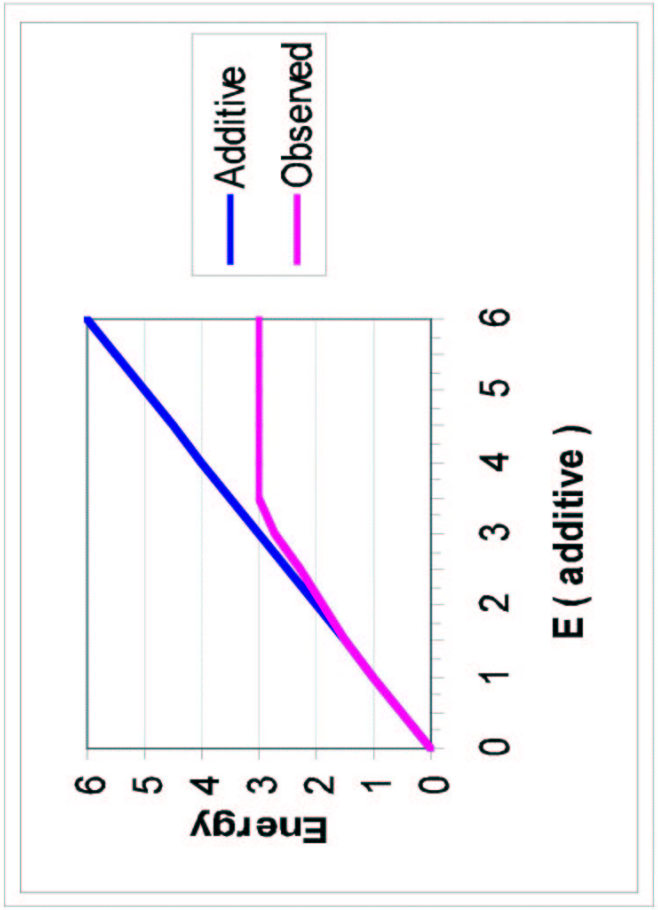
Non-independence of Mnt repressor-operator Interaction determine by a new quantitative multiple Fluorescence relative affinity (QuMFRA) assay

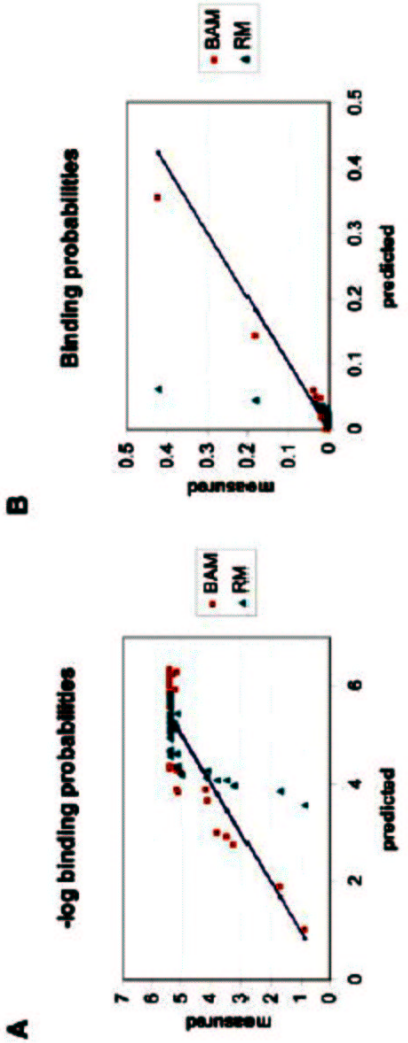
Tsz-Kwong Man and Gary D. Stormo
Nucleic Acids Research, 2001, Vol. 29, 2471-2478

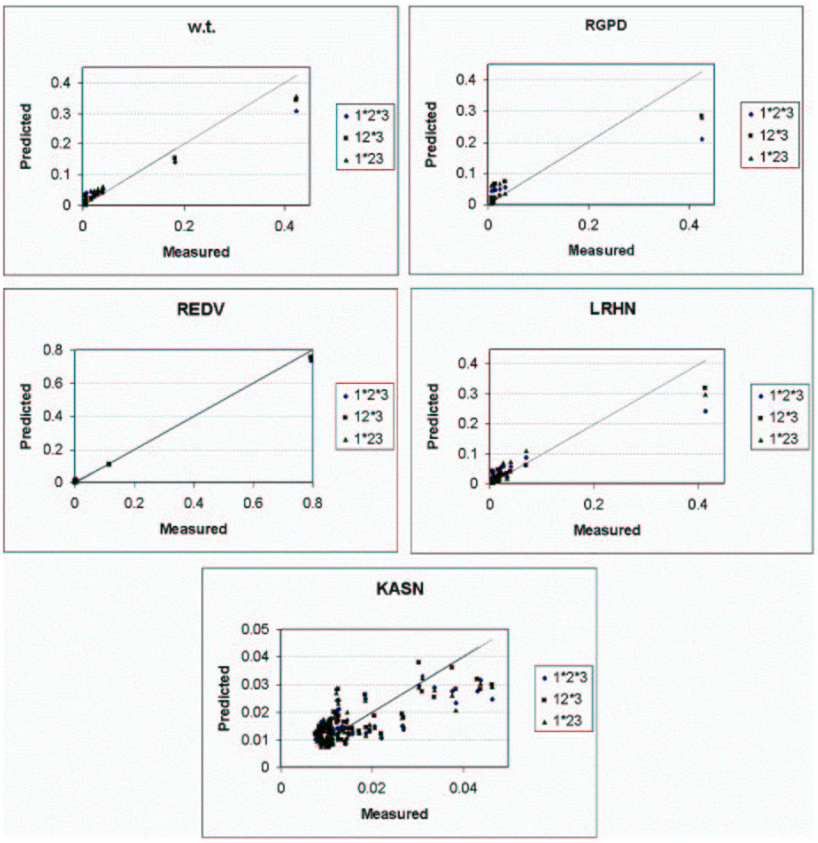



Mnt
1+2
measured
Mnt
Probability
Ka
1+2
M A G C A C G G A G T A T G


5'
3'
Arg-1
Glu 3
Arg 6
Asp 2
F1
F2
F3


Additive
Observed
Energy
0 1 2 3 4 5 6
E (additive)


A
-log binding probabilities
measured
predicted
BAM
RM
B
Binding probabilities
measured
predicted
BAM
RM



Correlation coefficients			
Zif268 variant	1*2*3	12*3	1*23
w.t.	0.973	0.986	0.987
RGPD	0.883	0.942	0.941
REDV	0.999	0.999	0.999
LRHN	**0.927**	0.978	0.956
KASN	0.695	**0.791**	**0.718**

doi:10.1016/S0022-2836(02)00917-8 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 323, 701–727

JMB

Probabilistic Code for DNA Recognition by Proteins of the EGR Family

Panayiotis V. Benos[1], Alan S. Lapedes[2] and Gary D. Stormo[1]*

Statistical

Algorithm for

Modeling

Interaction

Energies

Statistical Algorithm for Modeling Interaction Energies

Assuming equilibrium, Boltzmann leads us...

$$P(\underline{\mathbf{c}}) = \frac{e^{-E(\underline{\mathbf{c}})}}{\sum_{\underline{\mathbf{x}}} e^{-E(\underline{\mathbf{x}})}}$$

Two kinds of *in vitro* selection experiments

- SELEX
  - Fixed protein sequence selects preferred DNA sites from a random pool; usually returns more than one DNA sequence for each protein
- Phage display
  - Fixed DNA sequence selects preferred protein from a random pool (only "critical residues" are randomized); usually returns more than one protein sequence

SAMIE

In a SELEX experiment:

$$P(\underline{\mathbf{n}} \mid \underline{\mathbf{A}}) = \frac{P_{ref}(\underline{\mathbf{n}})\ e^{-E(\underline{\mathbf{n}}, \underline{\mathbf{A}})}}{\sum_{\underline{\mathbf{x}}} P_{ref}(\underline{\mathbf{x}})\ e^{-E(\underline{\mathbf{x}}, \underline{\mathbf{A}})}}$$

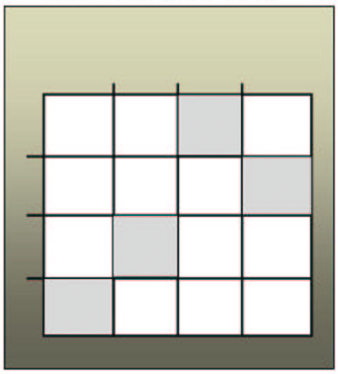
where energy:

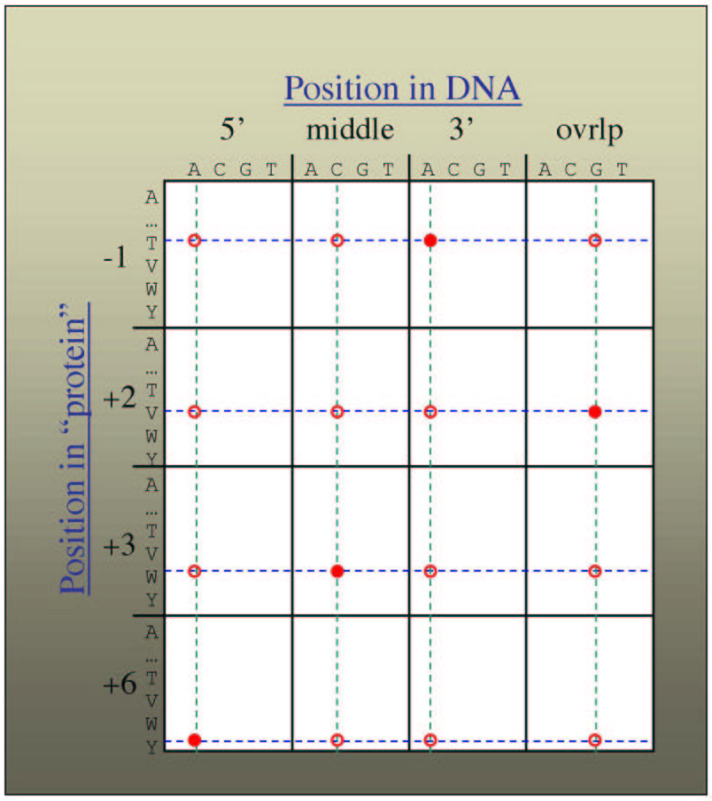
$$E(\underline{\mathbf{n}}, \underline{\mathbf{A}}) = \sum_{ijab} C_{ij}\ n_i^a\ T_{ij}^{ab}\ A_j^b$$

$$E(\underline{\mathbf{n}}, \underline{\mathbf{A}}) = \sum_{ijab} C_{ij}\ n_i^a\ T_{ij}^{ab}\ A_j^b$$

TVWY selects ***ACAG***

"*Connectivity matrix* "C





SAMIE

Training algorithm.

- Calculate T_{ij}^{ab} that maximises log-likelihood of the data.
- *Steepest ascents* method.

$$P(\{data\}) = \Pi_{\{data\} \ni \underline{n}} P(\underline{n} \mid \underline{A})$$

$$\log P(\{data\}) = \Sigma_{\{data\} \ni \underline{n}} \log P(\underline{n} \mid \underline{A})$$

$$P(\underline{n} \mid \underline{A}) = \frac{P_{ref}(\underline{n})\ e^{-E(\underline{n}, \underline{A})}}{\Sigma_{\underline{x}} P_{ref}(\underline{x})\ e^{-E(\underline{x}, \underline{A})}}$$

The EGR data

Binding examples (*three fingers*).

- 1033 examples of EGR protein family (919 non-redundant)
  - 322(304) SELEX and 431(399) phage display.

Randomisation experiments (*single fingers*).

- 366 distinct SELEX results
  - 70 tetra-nucleotides were selected by 180 tetra-peptides.
- 444 distinct phage display results
  - 361 tetra-peptides were selected by 59 tetra- nucleotides.

Total of 2009 base-amino acids interacting pairs

Table 2. The energy matrix as it was calculated by SAMIE when trained on the COMBINED_6 dataset

	Finger position = −2; base position = 3				Finger position = +2; base position = 4				Finger position = +3; base position = 2				Finger position = +6; base position = 1			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	1.29	0.92	1.32	6.42	0.25	−0.87	1.32	0.58	0.54	2.12	2.27	−1.14	−0.33	1.38	−0.13	0.72
C	5.21	4.71	6.29	4.51	3.86	−0.93	6.92	4.94	4.16	1.89	6.27	−0.06	4.68	5.07	5.11	5.39
D	1.93	−2.36	0.53	−0.68	2.27	0.42	−0.44	−0.29	5.82	−1.32	7.20	6.32	1.75	0.64	1.17	−0.21
E	0.34	−1.06	−0.05	−0.96	6.34	−1.72	1.68	0.29	1.30	0.43	1.96	1.16	0.09	1.68	1.26	−0.06
F	5.21	4.71	1.33	4.51	3.86	5.55	1.36	4.94	4.16	6.92	6.27	5.13	4.68	0.42	5.11	5.39
G	2.45	0.44	1.07	0.47	0.23	−1.56	0.39	−0.55	0.27	1.43	2.27	−0.51	2.62	2.62	1.61	1.54
H	1.41	−0.10	1.42	−0.65	−0.36	−0.99	0.20	−1.76	0.26	2.64	−1.31	6.72	1.73	6.10	−0.06	0.33
I	1.55	1.56	1.36	1.77	1.05	5.55	6.45	5.93	5.68	2.56	7.23	1.23	1.61	1.77	5.76	6.46
K	0.93	−0.57	−1.04	−1.14	6.59	−0.84	2.30	6.47	5.54	7.47	0.77	1.04	0.67	6.59	−1.01	−1.36
L	1.16	0.79	7.89	−0.46	7.05	6.63	3.50	1.34	6.93	1.54	8.16	0.89	1.09	2.44	0.79	7.94
M	6.52	4.59	6.35	−1.55	6.33	5.55	2.41	0.26	0.26	7.32	2.06	−0.29	6.91	5.71	5.46	6.27
N	0.94	0.73	0.06	−1.33	1.14	−1.72	0.76	1.21	−3.73	0.86	−0.005	−0.80	−0.61	0.48	−1.00	1.75
P	7.33	5.89	7.36	0.75	7.01	0.17	2.27	6.58	6.20	8.16	2.85	1.73	7.70	2.38	0.21	0.74
Q	−2.50	6.32	0.32	−0.52	0.20	−1.32	1.76	0.30	−0.26	1.34	0.77	0.35	−0.92	6.10	−0.06	0.33
R	1.01	3.26	−1.36	0.46	1.66	−1.05	2.11	7.37	6.59	3.54	2.56	2.14	0.36	0.51	−3.33	1.29
S	1.06	0.73	0.61	−0.57	2.29	−1.45	0.26	−0.91	6.06	1.15	1.69	−1.65	−0.21	0.08	−0.40	−0.39
T	0.27	1.32	0.59	−1.33	−0.18	−1.52	1.29	1.23	7.50	−0.04	3.26	0.26	−1.19	0.20	−0.68	−0.46
V	7.59	6.30	1.57	0.92	7.01	0.17	1.76	0.57	8.01	0.79	8.52	0.80	1.13	1.98	0.46	0.76
W	5.21	4.71	6.29	4.51	3.86	−0.52	0.67	−0.71	4.16	6.92	6.27	−0.06	4.68	5.07	5.11	−0.27
Y	5.21	4.71	−0.05	4.51	4.95	−1.25	1.43	5.31	4.16	6.92	1.27	5.13	−1.37	5.48	5.28	5.61

The COMBINED_6 dataset consists of both SELEX and phage display data, according to the many-to-one model of interactions. The energy values presented here have been normalized so that the average binding constant for each contact is 1.0. This normalization does not affect the calculation of probabilities, as described in the text. The normalized matrix shows that the energetic potential of a base–amino acid contact varies, depending on its position in the DNA target and the protein.

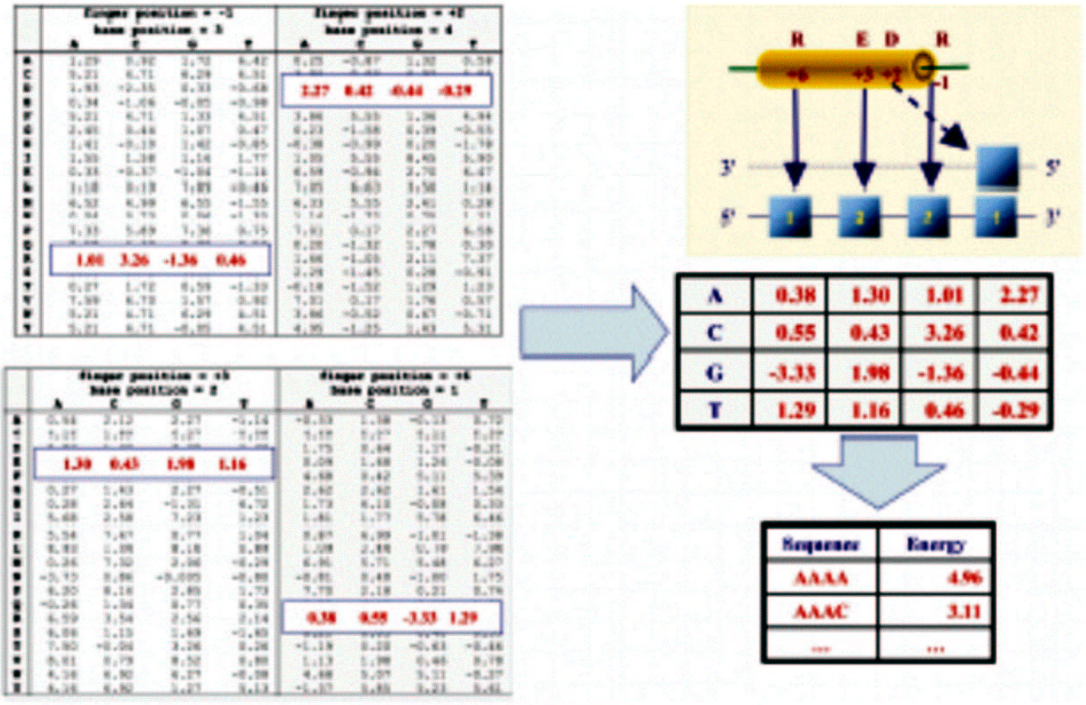
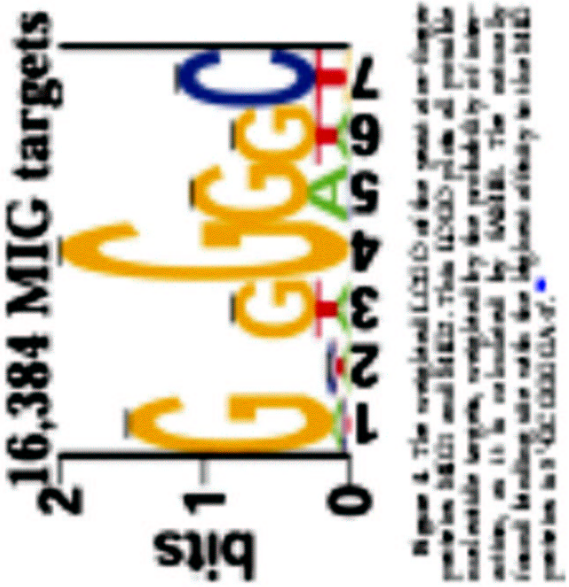
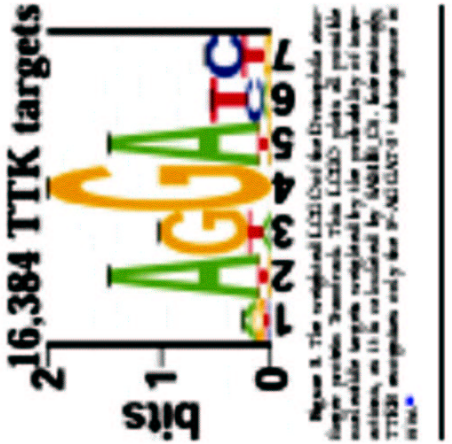
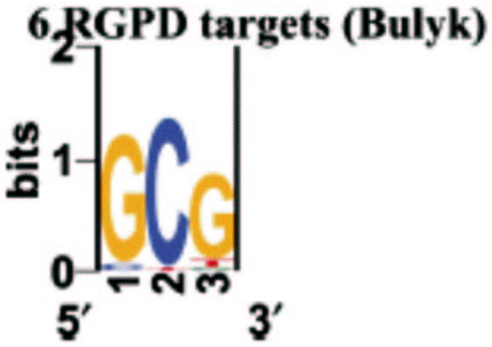
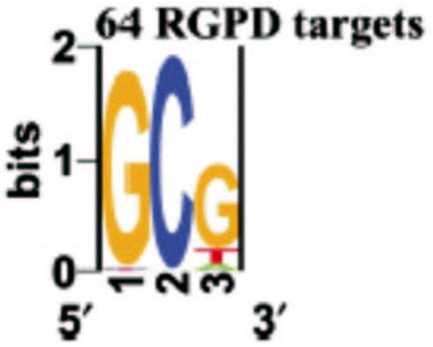


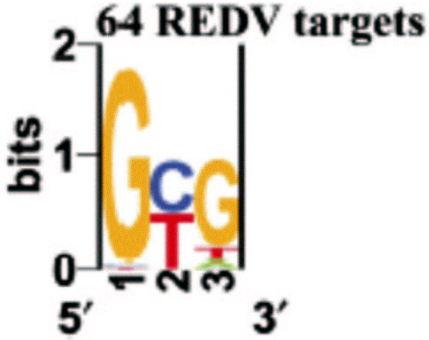
Figure 9 (legend opposite)


2
1
0
bits
16,384 MIG targets
1 2 3 4 5 6 7


2
1
0
bits
16,384 TTK targets
1 2 3 4 5 6 7





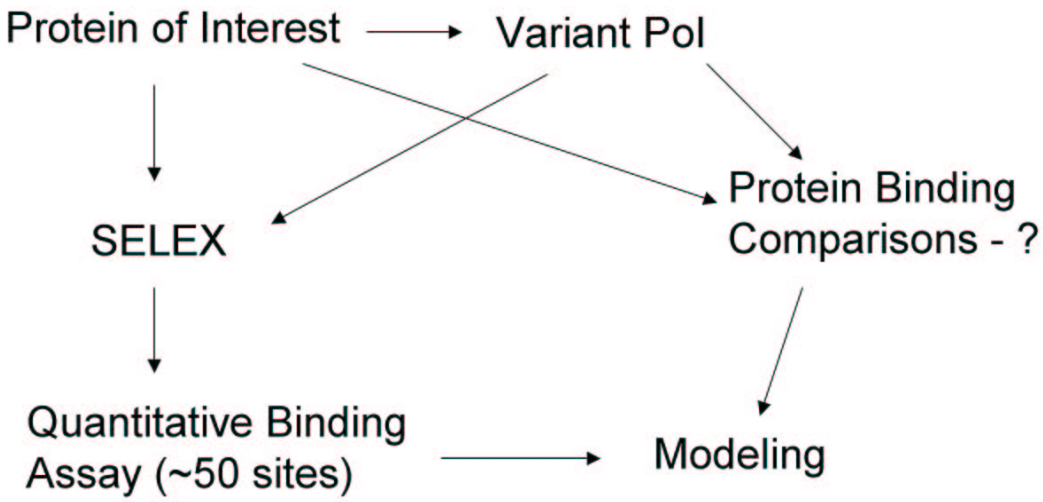


Quantitative Predictions

Source/protein	Correlation Coefficient
Hamilton	
EGR1	0.75
WT1	0.76
Segal (several variants)	0.82
Miller	
EGR1	0.93
D20A	0.81
Bulyk	
EGR1	0.61
RGPD	0.99
REDV	0.73
LRHN	0.56
KASN	0.16

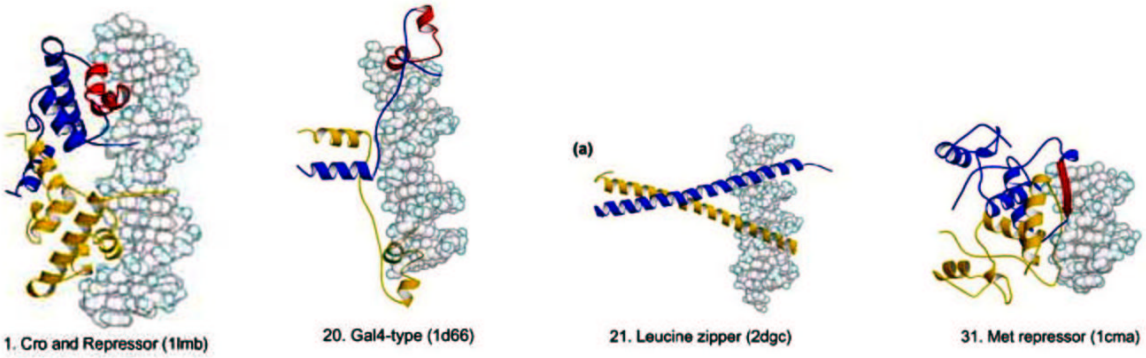

Alan Lapedes

Next: need more quantitative data for protein families of interest



Being done by Post-Docs: JJ Liu, Yong Yin

Examples of DNA-binding Proteins



HTH ZnF Leu zipper ribbon

from Luscombe *et al.* (2000)
http://genomebiology.com/2000/1/1/reviews/001.1