

Computational Molecular Engineering Nucleic Acid Binding Proteins and Enzymes

by

Faisal Reza

Department of Biomedical Engineering
Duke University

Date: _____

Approved:

Bruce R. Donald, Co-Chair

Jingdong Tian, Co-Chair

Thomas H. LaBean

Kam W. Leong

William M. Reichert

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the Department of Biomedical Engineering
in the Graduate School
of Duke University

2010

ABSTRACT

Computational Molecular Engineering
Nucleic Acid Binding Proteins and Enzymes

by

Faisal Reza

Department of Biomedical Engineering
Duke University

Date: _____

Approved:

Bruce R. Donald, Co-Chair

Jingdong Tian, Co-Chair

Thomas H. LaBean

Kam W. Leong

William M. Reichert

An abstract of a dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the Department of Biomedical Engineering
in the Graduate School
of Duke University

2010

Copyright © 2010 by Faisal Reza
All rights reserved

Abstract

Interactions between nucleic acid substrates and the proteins and enzymes that bind and catalyze them are ubiquitous and essential for reading, writing, replicating, repairing, and regulating the genomic code by the proteomic machinery. In this dissertation, computational molecular engineering furthered the elucidation of spatial-temporal interactions of natural nucleic acid binding proteins and enzymes and the creation of synthetic counterparts with structure-function interactions at predictive proficiency. We examined spatial-temporal interactions to study how natural proteins can process signals and substrates. The signals, propagated by spatial interactions between genes and proteins, can encode and decode information in the temporal domain. Natural proteins evolved through facilitating signaling, limiting crosstalk, and overcoming noise locally and globally. Findings indicate that fidelity and speed of frequency signal transmission in cellular noise was coordinated by a critical frequency, beyond which interactions may degrade or fail. The substrates, bound to their corresponding proteins, present structural information that is precisely recognized and acted upon in the spatial domain. Natural proteins evolved by coordinating substrate features with their own. Findings highlight the importance of accurate structural modeling. We explored structure-function

interactions to study how synthetic proteins can complex with substrates. These complexes, composed of nucleic acid containing substrates and amino acid containing enzymes, can recognize and catalyze information in the spatial and temporal domains. Natural proteins evolved by balancing stability, solubility, substrate affinity, specificity, and catalytic activity. Accurate computational modeling of mutants with desirable properties for nucleic acids while maintaining such balances extended molecular redesign approaches. Findings demonstrate that binding and catalyzing proteins redesigned by single-conformation and multiple-conformation approaches maintained this balance to function, often as well as or better than those found in nature. We enabled access to computational molecular engineering of these interactions through open-source practices. We examined the applications and issues of engineering nucleic acid binding proteins and enzymes for nanotechnology, therapeutics, and in the ethical, legal, and social dimensions. Findings suggest that these access and applications can make engineering biology more widely adopted, easier, more effective, and safer.

keywords: molecular engineering, computational biology, synthetic biology, protein design, nucleic acid, binding protein



To my mother and father,
Khurshid Ara Begum and Rezaul Kabir

Contents

Abstract	iv
Contents	vii
List of Tables	xi
List of Figures	xii
List of Abbreviations and Symbols.....	xvi
Acknowledgements	xxi
1. Introduction.....	1
1.1. Intelligent design of evolved molecules	1
1.2. Organization of this dissertation	5
2. Engineering molecules through computation	11
2.1. Motivation	11
2.2. Natural systems.....	12
2.2.1. DNA	12
2.2.2. Nucleic acid binding proteins and restriction endonucleases	15
2.3. Modeling natural systems	23
2.3.1. Gene-protein circuit modeling	23
2.3.2. DNA-enzyme structure modeling.....	33
2.4. Engineering of models	40
2.4.1. Operating philosophy.....	43
2.4.2. Input model structure visualization and preparation.....	46
2.4.3. Rotamer library.....	49

2.4.4. Energy function.....	50
2.5. Discussion	52
3. Single-conformation engineering of nucleic acid binding proteins	53
3.1. Motivation	54
3.2. Overview	55
3.2.1. Computational filtering approaches.....	56
3.2.2. Biological focusing approaches	62
3.2.3. Coupling computational and biological approaches: CF-BF	64
3.3. Methods.....	64
3.3.1. Molecular system selection.....	64
3.3.2. Primary structure BF	68
3.3.3. Secondary structure BF	69
3.3.4. Tertiary and quaternary structure BF.....	69
3.3.5. Engineering validations.....	71
3.4. Results.....	72
3.4.1. Objective of CF-BF.....	72
3.4.2. Primary Structure (PS) Biological Focusing.....	74
3.4.3. Secondary Structure (SS) and Hydropathy BF	77
3.4.4. Tertiary (TS) and Quaternary (QS) Structure BF.....	79
3.4.5. Computational validation via structural mutagenesis	81
3.4.6. Preliminary experimental validation.....	83
3.5. Discussion	85

4. Multiple-conformation engineering of nucleic acid binding proteins	90
4.1. Motivation	91
4.2. Overview	91
4.3. Methods.....	94
4.3.1. Scanning NABP residues for mutation tolerance.....	94
4.3.2. Redesigning NABP residues for functional mutations.....	95
4.3.3. Engineering validations.....	95
4.4. Results.....	96
4.4.1. Molecular engineering protocol for NABPs.....	97
4.4.2. Scanning for mutation tolerance and experimental validation ...	112
4.4.3. Computational redesign for functional S81 mutants and experimental validation.....	117
4.5. Discussion	131
5. Open-source molecular engineering	138
5.1. Motivation	138
5.2. Overview	139
5.3. Modeling flexibility in nucleic acid binding protein redesign.....	140
5.4. Computational redesign in practice	141
5.4.1. Scanning.....	143
5.4.2. Redesign	143
5.4.3. Native structure recovery	143
5.4.4. Molecular rebuilding.....	147
5.5. Discussion	148

6. Applications and issues of molecular engineering.....	149
6.1. Motivation	150
6.2. Bionanotechnologies	151
6.3. Therapeutics	159
6.4. Ethical, legal, and social issues (ELSI).....	163
6.5. Discussion	165
7. Conclusion	166
7.1. Contributions of this dissertation	166
7.2. Intelligent evolution of molecular design	171
Appendix.....	173
A.1. Amino acid templates.....	173
A.2. Amino acid rotamer library	178
A.3. Amino acid rotamer volumes	182
A.4. Energy function.....	183
A.5. Nucleic acid templates	184
References	185
Biography	225
B.1. Personal.....	225
B.2. Training.....	225
B.3. Research	225
B.4. Teaching and service.....	226
B.5. Awards and honors	227

List of Tables

Table 2-1: Sequence specific proteins that recognize DNA	16
Table 2-2: Computational and experimental parallels among DNA and restriction endonucleases	19
Table 2-3: Restriction endonuclease crystallographic structures	35
Table 2-4: R.PvuII crystallographic structures	37
Table 2-5: Amino acid pKa values for modeling titratable moieties.....	48
Table 3-1: Comparative protein candidate structures.....	67
Table 3-2: Origins and attributes of chosen proteins	75
Table 4-1: Computed ensemble-based binding affinities, rather than global minimum energy conformation-based bound energies, are better predictors of experimental outcomes	132
Supplementary Table 4-1: Pruning efficiency of redesigned R.PvuII S81 mutants for cognate and non-cognate DNA substrates.....	115

List of Figures

Figure 2-1: Principles of DNA	14
Figure 2-2: Principles of restriction endonucleases.....	17
Figure 2-3: Restriction endonuclease functional activity upon cognate DNA substrate.....	20
Figure 2-4: Restriction endonuclease binding and catalytic kinetics with cognate and non-cognate DNA substrates	21
Figure 2-5: Reaction coordinates of protein-DNA binding and catalysis	22
Figure 2-6: Single stage gene-protein circuit model.....	24
Figure 2-7: An oscillatory input signal can generate an output signal with oscillations compounded with noise	26
Figure 2-8: From analytical decomposition, the amplitude of output oscillations decreased with f_{in} , while the critical frequency, f_c , was calculated as the intersection between the “average noise level” curve and the “oscillation amplitude” curve.....	30
Figure 2-9: From numerical stochastic Gillespie simulation and FFT for dominant frequency, a critical frequency, f_c , consistent with the analytical decomposition was observed.....	31
Figure 2-10: The percentage of cells that generated f_{out} with fidelity, i.e. equal to f_{in} , reduced sharply at and beyond the critical frequency, f_c	32
Figure 2-11: Fixed basepair origin, F , for all unique arrangements of single basepairs.....	38
Figure 2-12: Bound cognate DNA substrate crystallographic structure exhibits deformation compared to cognate B-DNA model	39
Figure 2-13: Degree of DNA substrate deformation differs among REase-bound crystallographic structures.....	41
Figure 2-14: Operating philosophy of synthetic biomolecules inspired by modeling and engineering central dogma of natural biomolecules	45

Figure 3-1: Protein interfacial pocket engineering possibilities	57
Figure 3-2: CF-BF reduces the search space and the corresponding cost required to locate the global minimum energy conformation	65
Figure 3-3: Flowchart of coupled CF-BF criterion to engineer an IP	73
Figure 3-4: PS properties of chosen proteins	76
Figure 3-5: Tertiary and quaternary structure properties of remaining proteins after primary and secondary structure focusing.....	80
Figure 3-6: BF for putative engineered IP on original R.PvuII scaffold to bind R.EcoRV 5'-gatatc-3' substrate	82
Figure 3-7: Computational validation based on hydrogen bond and polar contacts with respect to steric hindrance patterns.....	84
Figure 4-1: Molecular engineering protocol for computing ensemble-based binding energies and affinities of protein-nucleic acids interactions	99
Figure 4-2: Input structure model of R.PvuII with interactions and residues of interest.....	101
Figure 4-3: Computational and experimental alanine scans of selected residues in R.PvuII	113
Figure 4-4: Computed binding affinities of redesigned R.PvuII S81 mutants to cognate DNA substrate.....	118
Figure 4-5: Enzymatic activities of redesigned R.PvuII S81 mutants.....	128
Figure 4-6: Modeling of global minimum energy conformations near the -3c::+3g BP of cognate DNA substrate 5'-cagctg-3' illustrates packing of redesigned R.PvuII S81 mutants	134
Figure 5-1: Open-source software engineering for computational redesign of nucleic acid binding proteins	142
Figure 5-2: Native structure recovery of residues in R.PvuII structure.....	144
Figure 5-3: Molecular rebuilding of R.PvuII placeholder residue A94 to redesigned wildtype residue A94Y	146

Figure 6-1: Wildtype R.PvuII-WT and redesigned R.PvuII-S81T and R.PvuII-S81A mutants in restriction-modification gene-protein circuits	157
Figure 6-2: Nucleic acid binding proteins and enzymes as gene reprogramming therapeutics.....	162
Supplementary Figure 3-1: Traditional dead-end elimination (DEE) criterion for nucleic acid binding proteins	61
Supplementary Figure 3-2: Secondary structure and hydrophathy properties of remaining proteins after primary structure focusing	78
Supplementary Figure 3-3: Synthetic R.PvuII gene validated by sequencing	86
Supplementary Figure 3-4: <i>In vivo</i> cell survival assay for R.PvuII Putative Engineered IP mutant in <i>E. coli</i>	87
Supplementary Figure 3-5: <i>In vitro</i> enzymatic function assay for R.PvuII Putative Engineered IP mutant in <i>E. coli</i>	88
Supplementary Figure 4-1: Experimental protocol for validating mutant proteins.....	100
Supplementary Figure 4-2: minimized side-chain Dead-End Elimination (minDEE) criterion for nucleic acid binding proteins	108
Supplementary Figure 4-3: Redesigned R.PvuII S81 mutants computed binding affinities to the cognate and non-cognate DNA substrates	119
Supplementary Figure 4-4: Enzyme synthesis and mutagenesis.....	123
Supplementary Figure 4-5: DNA sequencing chromatograms validates gene sequence of scanned and redesigned R.PvuII S81 mutants.....	124
Supplementary Figure 4-6: Design and evaluation of DNA substrate for REases	126
Supplementary Figure 4-7: DNA sequencing chromatograms validates restriction sequence of +cut DNA, i.e. largest fragment, from DNA substrate after complete digestion by redesigned R.PvuII S81 mutants.....	127


List of Abbreviations and Symbols

1D	one dimensional
3D	three dimensional
Å	Angstrom (1 Å = 10 ⁻¹⁰ meters)
a	adenine
A, Ala	alanine
AA	amino acid
AS	active site
ATP	adenosine triphosphate
B, Asx	aspartic acid or asparagine (ambiguous)
BB	backbone
R.BamHI	REase genus and species <i>Bacillus amyloliquefaciens</i> strain H 1 st in order identified
<i>bamHIR</i>	gene for R.BamHI
R.BglIII	REase genus and species <i>Bacillus globigii</i> 2 nd in order identified
<i>bglIIR</i>	gene for R.BglIII
BP	base pair (of DS DNA)
c	cytosine
C, Cys	cysteine
D, Asp	aspartic acid (carboxylate anion is aspartate)
Da	Dalton (= 1 atomic mass unit)
DBP	DNA binding protein, a type of NABP
DEE	dead-end elimination
DIY	do-it-yourself
DNA	deoxyribose nucleic acid

DS	double stranded
DSB	double stranded break
E, Glu	glutamic acid (carboxylate anion is glutamate)
R.EcoRI	REase genus and species <i>Escherichia coli</i> strain RY13 1 st in order identified
<i>ecoRIR</i>	gene for R.RcoRI
R.EcoRV	REase genus and species <i>Escherichia coli</i> strain RY13 5 th in order identified
<i>ecoRVR</i>	gene for R.EcoRV
EDTA	ethylenediaminetetraacetic acid
F, Phe	phenylalanine
FFT	fast Fourier transform
g	guanine
G, Gly	glycine
GMEC	global minimum energy conformation
H, His	histidine
I, Ile	isoleucine
ID	identification
indel	insertion or deletion
IPTG	isopropyl- β -D-thiogalactoside
J, Xle	leucine or isoleucine (ambiguous)
K, Lys	lysine
K*	ensemble-based computed approximation to K_a
K_a	association constant
K_d	disassociation constant
k_B	Boltzmann constant
k_{cat}	first-order rate constant

L, Leu	leucine
M, Met	methionine
mRNA	messenger RNA
MTase	methyltransferase
n	a, c, g, or t
N, Asn	asparagines
N/A	not applicable
NA	nucleic acid
NABP	nucleic acid binding protein (e.g. DBP or RBP)
NDB	Nucleic Acid Database
NHEJ	non-homologous end joining
NT	nucleotide
ORF	open reading frame
P, Pro	proline
PCR	polymerase chain reaction
PDB	Protein Data Bank
pKa	acid disassociation constant, $-\log_{10}(K_a)$
PS	primary structure
M.PvuII	MTase genus and species <i>Proteus vulgaris</i> 2 nd in order identified
<i>pvuIIM</i>	gene for M.PvuII
r	a or g
R.PvuII	REase genus and species <i>Proteus vulgaris</i> 2 nd in order identified
<i>pvuIIR</i>	gene for R.PvuII
Q, Gln	glutamine
QS	quaternary structure

R, Arg	arginine
R ⁴	residue replacement R-group refinement
RBP	RNA binding protein, a type of NABP
RBS	ribosomal binding site
REase	restriction endonuclease
R-M	restriction-modification
RMSD	root-mean-square deviation
RNA	ribose nucleic acid
RNAi	RNA interference
s	g or c
S, Ser	serine
SC	side-chain
SNP	single nucleotide polymorphism
SS	secondary structure
ss	single stranded
t	thymine
T, Thr	threonine
TAE	Tris/Acetate/EDTA
TBE	Tris/Borate/EDTA
TFO	triplex forming oligonucleotide
Tris	tris(hydroxymethyl)aminomethane
tRNA	transfer RNA
TS	tertiary structure
TU	template unit
V, Val	valine
vdW	van der Waals
w	a or t

W, Trp	tryptophan
WT	wildtype
X, Xaa	unspecified or unknown amino acid (ambiguous)
XFP	X fluorescent protein, X = G (green), Y (yellow) etc.
y	c or t
Y, Tyr	tyrosine
Z, Glx	glutamic acid or glutamine (ambiguous)
ZFP	zinc finger protein
φ	protein BB dihedral angle, by C_{i-1} -N-CA-C
ψ	protein BB dihedral angle, by N-CA--C- N_{i+1}
ω	protein BB dihedral angle, by CA-C- N_{i+1} + CA_{i+1}
χ^n	protein SC dihedral angle, e.g. χ_1 by N-CA-CB-CG
	anaglyph 3D image, viewable in 3D with red (over left eye)-blue (over right eye) glasses

Acknowledgements

*If I have seen a little further,
it is by standing on the shoulders of giants.*

– Isaac Newton
natural philosopher

This work is possible due to the people I have met and opportunities I have received. To those acknowledged here, and to those left unacknowledged, know that I am more grateful than the following can convey.

I thank my Dissertation Committee Co-Chairs, Bruce R. Donald and Jingdong Tian, for their exceptional advice. Bruce has shared remarkable intelligence and optimism in computational molecular design. Jingdong has shown steadfast enthusiasm and determination in engineering synthetic molecules. It has been a privilege having these dedicated advisors broaden and deepen my development as an independent researcher.

I thank my Dissertation Committee Members, Thom LaBean, Kam Leong, and William “Monty” Reichert, for their generous guidance. Thom has offered his creativity and assistance towards achieving my goals. Kam has shared his astute wisdom and outlook for the prospects of my work. Monty has instilled his clear insights, inquiries, and faith in my abilities. The confidence of these visionary professors has sustained and supported me through graduate school.

I thank co-workers, Syandan Chakraborty, Cheng-Yu Chen, Nicholas Christoforou, Ivelin Georgiev, Kuo-Sheng Ma, John MacMaster, Kyle Roberts, Cheemeng Tan, Qihai Wang, Lingchong You, Fan Yuan, and Peijun Zuo for sharing their enthusiasm for knowledge and the pursuit of discovery.

I thank members of Duke University's Department of Biomedical Engineering, like Kathy Barbour, Ashutosh Chilkoti, Ned Danieleley, Marcus Henderson, David Katz, Barry Myers, Ellen Ray, Susan Story-Hill, and George Truskey. At the Institute for Genome Sciences and Policy, among many I thank Huntington Willard. I thank those at the Pratt School of Engineering, including Kathleen Cahill, Marianne Hassan, Tom Katsouleas, and Carla Sturdivant. I appreciate the research support from the National Institutes of Health, Sigma Xi, The Scientific Research Society, Biomedical Engineering Society, Duke University's Center for Biomolecular and Tissue Engineering, and its Computational Biology and Bioinformatics Program.

I thank mentors at the Massachusetts Institute of Technology (MIT), George Church, Doug Lauffenburger, Rafael Reif, and Greg Stephanopoulos and teachers at the Bronx High School of Science, Mitch Fox, John Kelly, Sherrill Mirsky, and Joel Seidenstein, for fostering a passion for research.

Foremost, I thank my mother, Khurshid Ara Begum, and father, Rezaul Kabir, for their love, strength, and perseverance. They devoted their lives to my own. My successes are, and will always be, their success.

With sincere appreciation to those who have eased and enlivened
my time and way through Duke University, Durham, North Carolina, USA

Charles Anamelechi	BME	Endothelial cell-based grafts
Serkan Apaydin	CS	NMR structure-based assignment
Frances H. Arnold	CalTech	Directed evolution, academic career
Kathryn Ashley	BME	Payroll coordination
Jennifer Avery	IGSP	Posters
Nima Badie	BME, CBTE	Cardiomyocyte modeling
Jerome F. Baker	Sigma Xi	Society executive direction
Kathy Barbour	BME	Departmental support
David Becker	CS	SGE, MPI, rsh
Weining Bian	BME	Hydrogels for muscle tissue
Jeremy N. Block	BIOCHEM	KiNG, KinImmerse, GSS Comm.
John A. Board	CSEM	Computational support
I. Regina Borkoski	BMES	Student programs
Philip E. Bourne	UCSD, PDB	PDB format, structural bioinformatics
Kevin Bowen	Sigma Xi	Student Research Conference
Rachael Brady	CSEM	Visualization, anaglyphs
Eileen Brand	CBTE, CBIMMS	Graduate support, equipment
Matthew A. Brown	BME	Teaching lab equipment

Melissa Brown	BME, CBTE	Endothelial progenitor cells
August Burns	FIP	Poster printing
Kathleen Cahill	Pratt	Grants, reimbursements
Daniel J. Callahan	BME, CBIMMS	pH sensitive ELPs
Isabel Cardenas-Navia	BME	Mock study sections
Lynda M. Cecere	IGSP	Scheduling assistance
Syandan Chakraborty	BME	Physiology
Lih Mei and Jack Chao	Grace's Café	Chinese lunch specials, buffet
Cheng-Yu Chen	BIOCHEM	Kinetics assays, analysis
Ashutosh Chilkoti	BME, CBIMMS	ELPs, departmental support
Nicholas Christoforou	BME	ES/iPS cells
Erica Clayton	BME	Payroll
Jeffrey M. Coles	MEMS, CBIMMS	Articular joint tribology
Stephen L. Craig	CHEM	Computational chemistry
Thomas A. Darden	NIEHS	Force fields, molecular mechanics
Robert Cook-Deegan	IGSP	IP, USPTO, ELSI
Ned D. Danieleley	BME	IT, networking
Mark R. DeLong	IGSP	IT, DSCR queues
Michael DeSoto	BME	Mechanical tools
Bruce R. Donald	CS, BIOCHEM	minDEE, A*, K*, NMR, iGEM
Nelita T. Elliott	BME	Electric field-mediated gene delivery

Stuart Endo-Streeter	CS	Kinetic assays, labeling
Drew Endy	MIT, Stanford	Biological abstraction, iGEM
M. Judah Folkman	Harvard	Angiogenesis, publishing
Joyce Franklin	BME	Reimbursements
Terrence S. Furey	IGSP	Genome browser, cluster access
Pablo Gainza-Cirauqui	CS	K* optimization, bounds
Andres Garcia	MEMS	Tapping-mode AFM
Anthony R. Geonnotti	BME	Chalk Talks, modeling transport
Ivelin Georgiev	CS	minDEE, K*, GrsA-PheA
Michelle Gignac	SMiF	Biological SEM/TEM
Robert Gotwals	NCSSM	Computational chem. of whiskering
Michael R. Gustafson II	ECE	Duke Chapter of Tau Beta Pi
Myra J. Halpin	NCSSM	iGEM
Paulette Harmon	Sigma Xi	Student Research Conference
Alexander Hartemink	CS, IGSP	The Oracle
Marianne Hassan	Pratt	Research funds, support
Jeffrey J. Headd	CBB, SBB	Protein structure error correction
Marcus H. Henderson	BME	Equipment, teaching lab
Jared Heymann	CHEM	Biomaterials, Fe binding chemistry
Mengchi Ho	ENVIRON	Duke Chapter of Sigma Xi
Celeste Hodges	CS	Lab management, visitors

Brendan Hodkinson	BIO	Duke Chapter of Sigma Xi
Swati Jain	CBB	BWM, RNA
Nathan Jenness	CBTE, CBIMMS	Chalk Talk Comm., Janus particles
Yong Jiang	MEMS, CBIMMS	AFM imaging of DNA
Gary and Amy Kapral	PhD Posters	Poster printing
Thomas Katsouleas	Pratt, ECE	NAE Summit, career development
David F. Katz	BME	Transport phenomena
Kathy L. Kay	Pratt	NAE Summit poster session
Daniel Keedy	BIOCHEM	NAMD, binding proteins
Thomas B. Kepler	BIOSTAT	Computational immunology
Donghwan Kim	BME	Protein microarrayer
Minkyu Kim	MEMS, CBIMMS	AFM biomolecule pulling
Dianne G. Kindel	BME	Teaching lab administration
Robert D. Kirkton	BME	Cardiac physiology
Marius Kluenger	BME	Global health
John A. Knesel	Sigma Xi	Southeast region direction
Thomas F. Knight	MIT EECS	Synthetic biology, BioBricks
Valerie, Ira Kolmaister	BMES	Student programs
Heidi Koschwanez	BME	Glucose sensors
Karina Kulangara	BME	Surface topography on cells
Thomas H. LaBean	CS, CHEM, BME	Bionanotechnology

Bonnie E. Lai	BME	Transport in microbicide gels
Tod A. Laursen	Pratt, MEMS	Scholar-administrator views
Curtis J. Layton	CBB, SBB	Protein-protein interfaces
Anne A. Lazarides	MEMS	Nanoscience seminars
Tae Jun Lee	BME	Cell cycle circuits
Whasil Lee	MEMS, CBIMMS	AFM biomolecule pulling
Kam W. Leong	BME, CBTE	Gene/drug delivery
I-Chien Liao	BME	Electrospinning, nanofibers
Leping Li	NIEHS	Genetic element Markov models
Ryan H. Lilien	Univ. of Toronto	Protein design, medicine
E. Allan Lind	BA	Leadership
Phillippe Luedi	CBB	Genomic imprinting
Kuo-Sheng Ma	BME	Oligo. synthesis, COC, Ag-DNA
John MacMaster	CS	Protein purification
Piotr E. Marszalek	MEMS, CBIMMS	Scanning probe microscopy
Jeff Martin	CS	Symmetric protein structures
Kathy McLane	Sigma Xi	National services, membership
Megan Mobley	ENVIRON	Duke Chapter of Sigma Xi
Alexander Motten	BIO	Duke Chapter of Sigma Xi
Krista L. Moyle	GS	Dissertation check
Sayan Mukherjee	STAT, IGSP	Statistical bioinformatics

Barry S. Myers	BME	One minute manager
David Needham	MEMS, CBIMMS	Micromanipulation, EDUK
Amy Norstrud	BME	Reimbursements
Lori Norton	BME	Glucose probes
Matthew T. Novak	BME, CBTE	Gene arrays
Gregory Nusz	BME, CBTE	Nanoparticle plasmonics
Uwe Ohler	CBB	Gene regulation, seminars
Taylan Ozdere	BME	Bistable gene switches
Anand Pai	BME	Quorum sensing potential
John B. Pormann	CSEM	Scalable computing
Jiayuan Quan	BME	Codon optimization
Mindy Quigley	CSEM	Computational support
Mahir H. Rabbi	MEMS, CBIMMS	AFM construction
Srinath Rangarajan	BME	DNA extension, amplification
Ellen M. Ray	BME	Business assistance
William M. Reichert	BME	Wound healing, presentations
Randy Rettberg	MIT	Synthetic biology, iGEM
Caroline Rhim	BME, CBTE	Artificial muscle differentiation
David C. Richardson	BIOCHEM	Protein structures physical models
Jane S. Richardson	BIOCHEM	Repairing protein structure data
Kyle E. Roberts	CBB	DEE/K* improvements, CFTR

Richard J. Roberts	NEB	REBASE, REase PDBs
Shandra L. Robertson	IGSP	Genomes@Grandover, posters
Elaine Ruger Emory	CBTE	Equipment support
Ishtiaq Saaem	BME	Oligonucleotide synthesis
William H. Safley	Pratt	Webservers, wikis
Scott C. Schmidler	STAT	Macromolecular structure
Robert J. Schutte	BME	Cellular responses to biomaterials
Lori A. Setton	BME	Duke Chapter of BMES
Joe Shamblin	CS	Filesystem, MPI
Amy L. Sheck	NCSSM	iGEM
Haige Shen	CBB	Bayesian statistics
William J. Shamblin	CS	MPI, MPIJava
Andrew J. Simnick	BME, CBIMMS	ELPs in drug delivery
Susan Story-Hill	BME	Scheduling assistance
Carla Sturdivant	CBTE, CBIMMS	Graduate support
Cheemeng Tan	BME	Gene circuit construction, f_c
Yu Tanouchi	BME	Noise in quorum sensing
Jingdong Tian	BME, IGSP	Gene synthesis, iGEM
David A. Tirrell	CalTech	Proteins as polymers
Vinalia Tjong	BME	Antibodies
Chittaranjan Tripathy	CS	Protein loop closure algorithms

George A. Truskey	BME	Departmental support, BMES
Dennis Tu	BME	Cell-cell communication
Alexei Valiaev	MEMS	CIERD
Gunjan Verma	CBB	Statistics, algorithms
Tuan Vo-Dinh	BME, FIP	Government labs view
Charles S. Wallace, Jr.	BME	Sheer stress
Mark D. Walters	SMiF	Advanced materials lab
Qihai Wang	BME	Protein expression, evaluation
Rui Wang	CBB	Statistics, modeling
Huntington F. Willard	IGSP	X inactivation, HACs
Ann H. Williams	Sigma Xi	Society organization
Susan Williford	GS	M.S., Ph.D. records, dissertation
Jo Rae Wright	GS	GSS Comm., grad. student policies
Mina Wu	BME, CBTE	Chalk Talk Comm., drug delivery
Anthony Yan	CS	NMR, MATLAB
Lingchong You	BME, IGSP	Gene-protein circuits, iGEM
Fan Yuan	BME, CBTE	Drug delivery, Kewaunee, iGEM
Stefan Zauscher	MEMS, CBIMMS	AFM, polymer science
David J. Zielinski	CSEM	DiVE, Virtools
Jianyang Zeng	CS	NMR assignment algorithms
Peijun Zuo	BME	Gene cloning

1. Introduction

*The known is finite,
the unknown infinite;
intellectually we stand on
an islet in the midst of an
illimitable ocean of inexplicability.
Our business in every generation
is to reclaim a little more land,
to add something to the
extent and solidity
of our possessions.*

– Thomas H. Huxley
British biologist

1.1. Intelligent design of evolved molecules

Molecules in nature evolved out of necessity or perished. Among these molecules, protein and enzymes evolved their functional prominence likely due to their greater diversity of constituent amino acids and broader effectiveness in catalytic capacities than those composed of ribose nucleic acid (RNA) during the dawn of the post-RNA world (1). In this world, a slightly different nucleic acid molecule from RNA, which was deoxyribose nucleic acid (DNA), carried biological information. In essence, interactions among DNA and the proteins that bind to them and the catalytic proteins, or enzymes, that facilitate a biochemical reaction with them, became paramount in nature for transforming biological information onto physiological function.

The importance of interactions between proteins and nucleic acids has not escaped our attention. The intelligent design from already evolved

nucleic acid binding proteins, which is termed redesign, presents significant advantages and notable caveats.

Among these advantages is the insight that evolved proteins have already been optimized along many properties, due to their having subjected and responded to various environmental cues and the considerable tests of time. The tests of time occurred both in the short-term lifetime of a functional protein in the cell, as well as the long-term lifetime of the protein in the cell population. For the short-term lifetime, the optimized properties included genesis during satisfactory translation off the ribosome, to proper folding from linear and semi-folded conformations, to reception of any requisite post-translational modifications, to stability in solution under perturbed cellular conditions, to binding, recognition, and possibly catalysis, of appropriate substrates, and demise during proteolytic degradation. In the long-term lifetime, these optimized properties, and the genes that encoded them, were subject to mutation and adaptation that determined the use and importance of the protein to the cell.

During redesign, it is prudent to utilize these natural advantages. A reasonable assumption is that redesigns of functional proteins are likely to also yield functional mutant counterparts that introduce or alter the desired property while, for the most part, maintaining the other existing properties. This assumption may be further supported if there is a relatively disparity in

size of the (large) molecule and of the (small) regions altered. Doing so could maintain overall stability and functionality, while the mutations restricted to the regions may confer the redesigned function.

Among the caveats is that depending on the degree of optimization, as well as type and amount of variation tolerated by the natural protein, perhaps even the slightest of mutations during redesign can instead upset an already-delicate balance among properties. This is often the case when the aforementioned disparity is quite small, so that each residue on average shares greater responsibility and interdependency in supporting the properties that have permitted the protein to remain fit and functional.

During design it would be necessary to observe these natural caveats. Thus, the aforementioned assumption can be amended to state that redesigns of functional proteins are likely to also yield functional mutant counterparts, if the mutations are compatible with the existing balance among properties or can introduce a new balance among properties that is suitable.

Proteins mutated in nature and in the laboratory are similar in some respects while differing in others. For example, the chemical constituencies of proteins, that are the amino acid residues that form polypeptide chains, are identical, regardless of whether they are in an organism (i.e. *in vivo*) or in a laboratory test tube (i.e. *in vitro*). Models of these proteins in a computer (i.e. *in silico*) may be further removed in apparent similarity, but through our

engineering efforts it is hoped that these representations reasonably describe the natural aspects. However, the setting for the protein, whether it be inside an organism or laboratory test tube, the environment that it affects and, in turn, affects it, and the motivation for mutating the protein highlight important differences for consideration, as follows.

In nature, it is entirely reasonable for the mutation of proteins, and the genes encoding them, to be driven by evolutionary instinct and shaped by the ensuing environmental pressures. Simply stated, the organism, or species for that matter, may not need to know how or why their proteins should mutate or remain unchanged with the times. Rather, the mutations would occur (or not), as determined by the necessities of regular and chance natural occurrences. Those mutations that happened to be favorable would permit the organism, and in turn the gene and protein within it, avoid obsolescence and extinction and continue being maintained and propagated. Thus, in nature, an organism's actions can be described as exploratory and its prerogative is to survive.

In the laboratory, it is often desirable for the mutation of proteins, and the genes encoding them, to be dictated by intelligent design approaches and observed under controlled conditions. Once again simply stated, we as researchers would be appreciative at discovering that mutant proteins functioned. Evolution can and has been mimicked in the laboratory in order

to achieve this initial goal by predicting the requisite combination of instinct and pressures necessary in order to mutate the protein with altered properties. Yet, as researchers we would also appreciate knowing why and how proteins should be mutated, ideally *a priori*, which is integral to the design process. Mutations would occur (or not) as determined by the intents of redesign. Those mutations that were evaluated to function as intended would permit us, as researchers, to understand and apply intelligent redesign principles across diverse genes, proteins, and conditions. Thus, in the laboratory, our actions can be deemed as hypothesis-driven and further our prerogative to understand and improve the world around us.

1.2. Organization of this dissertation

The organization of this dissertation presents our work on the intelligent design and redesign of evolved molecules, such as DNA and the proteins that bind them.

We believe that using computational means enables us model these molecules and form hypotheses *in silico* that can be then validated *in vitro* or *in vivo*.

We operate at the level of molecules and their constituent atoms in order to make accurate design and redesign mutation predictions *a priori*.

We put these beliefs and level of operation to the test through forward engineering from what we know to what we have yet to discover.

Taken together, we present this organization of the dissertation on computational molecular engineering nucleic acid binding proteins and enzymes, with respect to our publications and research meeting abstracts:

In Chapter 2, we discuss engineering molecules through computation. We first describe the features and activities of natural systems, such as DNA, DNA binding proteins, and restriction endonucleases (REases) in particular. We then discuss models of these natural DNA-protein interactions and perform two distinct studies, one dealing with these interactions across many nucleic acids and proteins in a gene-protein circuit, and another more focused on interactions of a single DNA substrate and protein at a structural level. We then posit that these models can be used not only to describe the existing activities of proteins and nucleic acids but engineer novel counterparts as well. Some of the material in Chapter 2 is based on a manuscript that was joint work with Cheemeng Tan and Lingchong You:

Tan C., Reza F., You L. Noise-limited frequency signal transmission in gene circuits. *Biophysical Journal*. 2007, 93: 3753-3761.

and partially from a research meeting abstract that was joint work with Cheemeng Tan, Lingchong You, Ivelin Georgiev, Bruce R. Donald, Jingdong Tian and Kuo-Sheng Ma:

Reza F. in collaboration with Tan C., You L. (modeling spatial-temporal interactions), Zuo P., Georgiev I., Donald B. R., Tian J. (engineering

structure-function interactions), Ma K-S. (nanoscale characterization of interactions). Modeling and engineering of nanomolecular interactions. *Sigma Xi, The Scientific Research Society 2008 Student Research Conference*. 2008, Washington, DC.

In Chapter 3, we discuss single-conformation engineering of nucleic acid binding proteins. This continues our investigation of engineering models of a single DNA substrate and protein at a structural level and on single conformations of these molecules. We examine computational filtering and biological focusing strategies and find that coupling them is advantageous. We apply a coupled filtering and focusing structure-based approach to altered the substrate specificity of a REase to another and perform preliminary experimental validations. Some of the material in Chapter 3 is based on a manuscript that was joint work with Peijun Zuo and Jingdong Tian:

Reza F., Zuo P., Tian J. Protein interfacial pocket engineering via coupled computational filtering and biological focusing criterion. *Annals of Biomedical Engineering: Special Issue: Systems Biology, Bioinformatics, and Computational Biology*. 2007, 35: 1026-1036.

and partially from a research meeting abstract that was joint work with Peijun Zuo and Jingdong Tian:

Reza F., Zuo P., Tian J. Theoretical and empirical perturbations of endonuclease-DNA biomolecular complexes. *Duke University Center for*

Biomolecular and Tissue Engineering Kewaunee Event. 2007, Durham, NC.

In Chapter 4, we discuss multiple-conformation engineering of nucleic acid binding proteins. This extends our continuing investigation of engineering models of a single DNA substrate and protein at a structural level to consider multiple conformations of these molecules simultaneously. Some of the material in Chapter 4 is based on a manuscript that was joint work with Qihai Wang, Ivelin Georgiev, Bruce R. Donald, and Jingdong Tian:

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Automated and accurate engineering of a superior nucleic acid enzyme. *In revision.*

and partially from a research meeting abstract that was joint work with Qihai Wang, Ivelin Georgiev, Bruce R. Donald, and Jingdong Tian:

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Molecular ensemble engineering and evaluation for targeted genome therapeutics. *Biomedical Engineering Society 2009 Annual Meeting. 2009, Pittsburgh, PA.*

In Chapter 5, we discuss open-source molecular engineering. We present the needs and benefits and address these by releasing and documenting our state-of-the-art suite of design algorithms for molecular engineering that model flexibility the redesign of proteins that act upon nucleic acids. Some of the material in Chapter 5 is based on a manuscript that was joint work with Ivelin Georgiev, Jingdong Tian, and Bruce R. Donald:

Reza F., Georgiev I., Tian J., Donald B. R. Open-source computational redesign of nucleic acid binding proteins. *To be submitted.*

and partially from a research meeting abstract that was joint work with Qihai Wang, Ivelin Georgiev, Bruce R. Donald, and Jingdong Tian:

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Computational and experimental scanning and redesign of nucleic acid proteins. *Sigma Xi, The Scientific Research Society 2009 Student Research Conference.* 2009, The Woodlands in Houston, TX.

In Chapter 6, we discuss applications and issues of molecular engineering. We consider the impact of such engineering in the areas of nanotechnologies, of therapeutics, and their ethical, legal, and social issues (ELSI). Some of the material in Chapter 6 is based on a manuscript that was joint work with Kuo-Sheng Ma, Ishtiaq Saaem and Jingdong Tian:

Ma K-S., Reza F., Saaem I., Tian J. Versatile surface functionalization of cyclic olefin copolymer (COC) with sputtered SiO₂ thin film for potential BioMEMS applications. *Journal of Materials Chemistry.* 2009, 19: 7914-7920.

partially from a research meeting abstract that was joint work with Jingdong Tian:

Reza F., Tian J. Engineering molecular interactions for targeted therapeutics and technologies. *National Academy of Engineering Grand Challenges National Summit.* 2009, Durham, NC.

and partially from a manuscript that was joint work with the Duke University Genetically Engineered Machines Program 2006:

Reza F., Chandran K., Feltz M., Heinz A., Josephs E., O'Brien P., Van Dyke B., Chung H., Indurkha S., Lakhani N., Lee J., Lin S., Tang N., LaBean T., You L., Yuan F., Tian J. Engineering novel synthetic biological systems. *IET Synthetic Biology*. 2007, 1: 48-52.

2. Engineering molecules through computation

The best way to have a good idea
is to have lots of ideas.

– Linus C. Pauling
American chemist

This chapter has been adapted partially from a manuscript that was joint work with Cheemeng Tan and Lingchong You:

Tan C., Reza F., You L. Noise-limited frequency signal transmission in gene circuits. *Biophysical Journal*. 2007, 93: 3753-3761.

and partially from a research meeting abstract that was joint work with Cheemeng Tan, Lingchong You, Ivelin Georgiev, Bruce R. Donald, Jingdong Tian, and Kuo-Sheng Ma:

Reza F. in collaboration with Tan C., You L. (modeling spatial-temporal interactions), Zuo P., Georgiev I., Donald B. R., Tian J. (engineering structure-function interactions), Ma K-S. (nanoscale characterization of interactions). Modeling and engineering of nanomolecular interactions. *Sigma Xi, The Scientific Research Society 2008 Student Research Conference*. 2008, Washington, DC.

2.1. Motivation

In order to engineer molecules through computation, it is useful to appreciate the properties of natural systems. Performing computational

modeling of these systems at various levels of abstraction can provide further insights that complement or inform experimental investigations. Engineering these molecules requires a healthy appreciation of the natural systems and a host of insights into the models of these systems.

2.2. Natural systems

The natural systems studied are deoxyribose nucleic acid (DNA) and restriction endonucleases (REases), a type of nucleic acid binding proteins (NABPs) with catalytic cleavage ability for specific sequences of DNA substrates. Interactions between DNA and REases are interesting and important for a number of reasons, possibly far too many to list. While these molecules have some features in common, such as both being biopolymers, they differ many ways, such as these polymers being composed of nucleic acids (NAs) in DNA and amino acids (AAs) in REases. Thus, before proceeding to modeling, it is useful to review the current understand of DNA and REases.

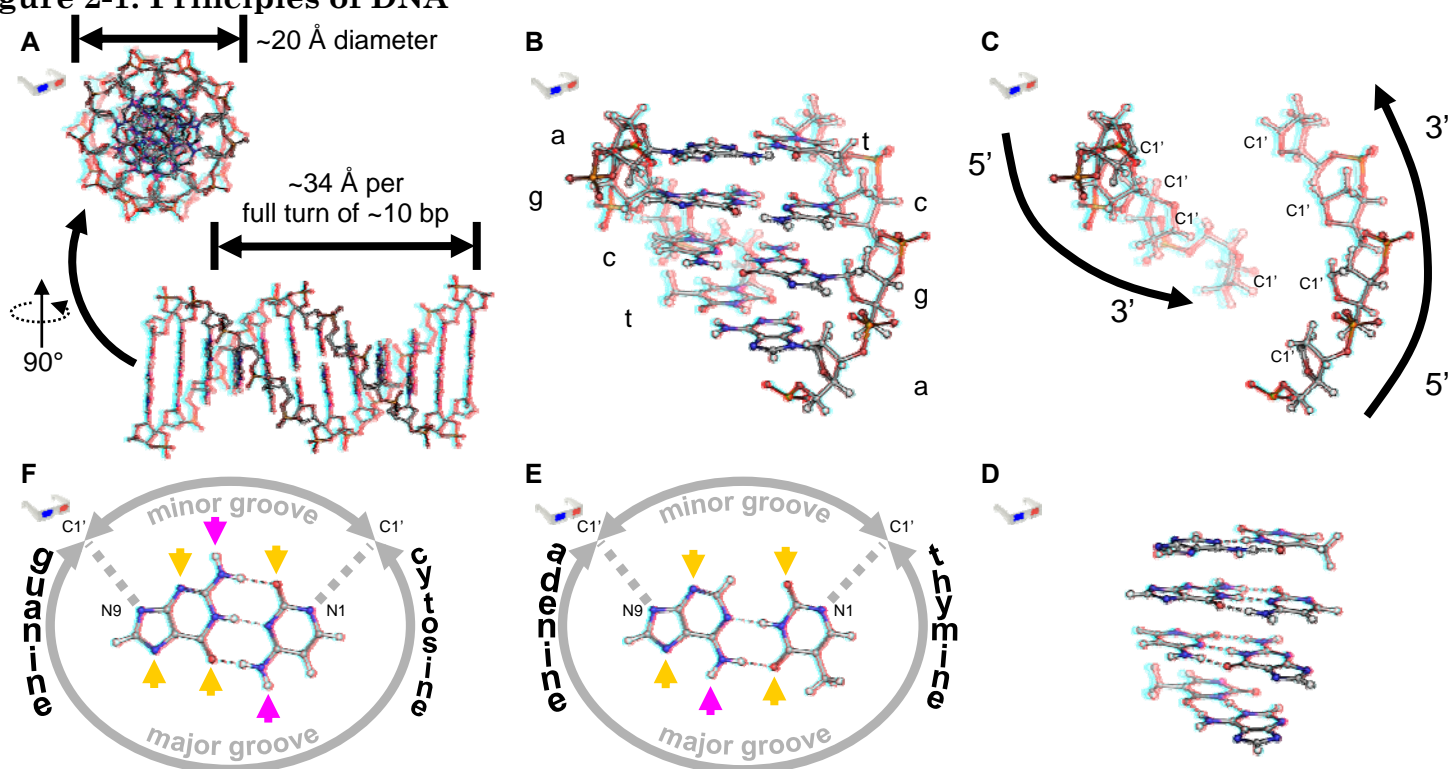
2.2.1. DNA

NAs, of which the predominant terrestrial types in biological systems are DNA and ribose nucleic acid (RNA), store and transmit genetic information. In 1944, Oswald and co-workers had successfully demonstrated this property of DNA to carry information (2). Less than a decade later, the three dimensional (3D) structure of DNA was correctly proposed by Watson

and Crick (3), for which they shared the Nobel Prize in Medicine or Physiology in 1962.

In the structure of DNA, both strands consisted of monomer units, or nucleotides (NTs), of a varying nitrogenous heterocyclic base covalently bonded to a form of deoxyribose sugar, which in turn are covalently tied together by phosphodiester bonds. These DNA basepairs (BPs) presented different spatial and biochemical profiles. The modeling of canonical B-DNA through fiber studies revealed further nuances in the overall polymeric structure (4). DNA contains four types of base, adenine (a), cytosine (c), guanine (g) and thymine (t). The axial and longitudinal aspects of this biopolymer enable interaction, and recognition, to occur at various scales and manners (5). From a distance, indirect recognition of DNA can occur through identification of the double helix and negative-charged phosphates spaced along the backbone (BB). Approaching closer, differences in the individual bases of DNA can be directly recognized (Figure 2-1). It is presumed that this indirect recognition is able to act on longer scales to localize a NABP onto the DNA. With the proximity conferred by localization, direct recognition is able to finely align a sequence specific NABP, such as a REase, to its intended DNA recognition site by making complementary polar and hydrogen bonding contacts between its interfacial (IP) amino acid

Figure 2-1: Principles of DNA



(A) DNA is nanoscale polymer of nucleic acids, (B) with each monomer containing one of four nucleotides, adenine, a, guanine, g, cytosine, c, or thymine, t. It is (C) composed of anti-parallel sugar phosphate double helices implicated in indirect recognition, (D) and of heterocyclic nitrogenous bases paired by two or three Watson-Crick hydrogen bonds. The N1 and N9 atom of (E) an a::t pair and (F) a g::c pair covalently bond to the C1' atoms of the backbone creating asymmetric major and minor grooves and hydrogen bond donor (magenta arrows) and acceptor (yellow arrows) atoms implicated in direct recognition.

residues and NTs found in the DNA's major or minor grooves (6).

In more recent times, DNA and its related molecules are being investigated in broader and more varied ways. By the end of the 20th century, all 3 billion BPs of the human genome were sequenced (7,8). The research continues, however, for understanding how the diverse array of NABPs encoded by these genomes in turn interact the DNA within these genomes. The work in this dissertation is a part of this ongoing pursuit.

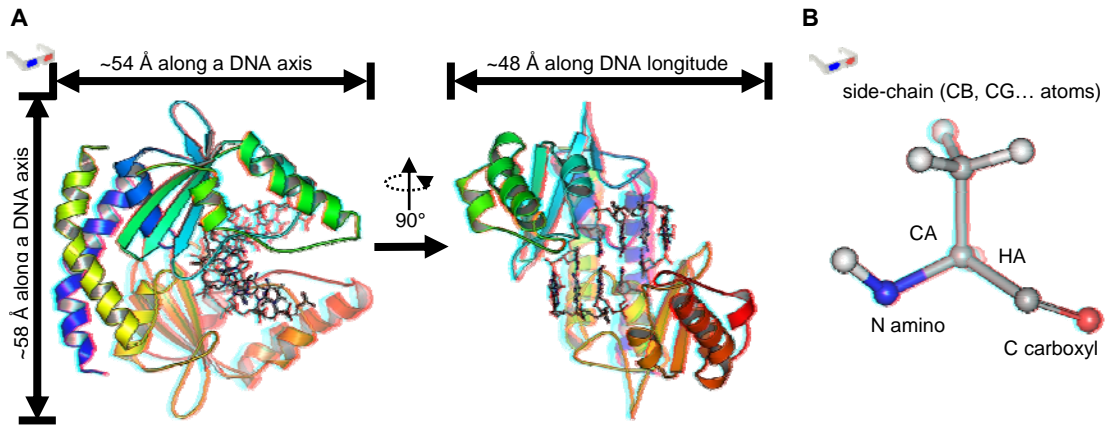
2.2.2. Nucleic acid binding proteins and restriction endonucleases

The cell has a wide variety of proteins and enzymes that bind and act upon NAs, such as transcription factors, polymerases, and ribosomes used in our work (Table 2-1). Among these enzymes of interest are REases due to their sequence-specific and robust catalytic function (Figure 2-2). Glimpses into the world of REases began in the 1960s. Arber and co-worker were studying “host controlled restriction of bacteriophages” and found that it provided bacteria with a defense mechanism against invading foreign DNA, such as viral DNA (9-11). Smith and co-workers purified the REases as well as the methyltransferase (MTase) from *Hemophilus influenzae*, characterized its properties, and deciphered the DNA BP sequence of its recognition site (12-15). Nathans and co-workers led in applying this “endonuclease R” (now

Table 2-1: Sequence specific proteins that recognize DNA

nucleic acid binding protein/enzyme	host(s) cell	substrate recognition site length	molecular function
alkaline phosphatase	archaea, bacteria, eukaryota	1 BP	hydrolase for dephosphorylation
endonuclease, homing	eukaryota	14-30 BPs, considerable tolerance of degeneracy	autonomous, self-catalyzing genetic elements
endonuclease, restriction	bacteria, archaea	4-8 BPs , limited tolerance of degeneracy	cleavage of foreign genome phosphodiester bonds
glycosylase	archaea, bacteria, eukaryota	1 BP	cleave nitrogenous base from nucleotide
histone (nucleosome core)	archaea, eukaryota	146 BPs	genetic "spool" for regulation of gene expression
kinase	archaea, bacteria, eukaryota	1 BP	transfer of phosphate group from ATP
ligase	archaea, bacteria, eukaryota	1 BP	formation of phosphodiester bond
methyltransferase	archaea, bacteria, eukaryota	3-8 BP	methylation of host genome nucleotides
repair, base excision	archaea, bacteria, eukaryota	1 BP	repair of oxidation, alkylation, hydrolysis, or deamination damage
repair, nucleotide excision	archaea, bacteria, eukaryota	2-30 BP	repair helix distorting and broader damage, e.g. via UV
repair, mismatch	archaea, bacteria, eukaryota	10-50 BP	repair error in replication and recombination mispairing
polymerase	archaea, bacteria, eukaryota	1 BP	polymerize deoxyribonucleotides into DNA strands
transcription factor	archaea, bacteria, eukaryota	dependent upon regulatory element(s)	controls transfer of information among nucleic acids, proteins

Figure 2-2: Principles of restriction endonucleases



(A) A REase is a nanoscale molecule composed of polymer peptide chains that recognizes and cleaves specific sequences of DNA, (B) with each monomer being one of twenty AAs. Each AA residue contains common backbone atoms composing its N-amino group, CA, HA, and C-carboxyl group, and a variable atoms composing its side chain or R-group. Here the R-group for the AA alanine, having CB, HB1, HB2, and HB3 methyl group atoms, is shown.

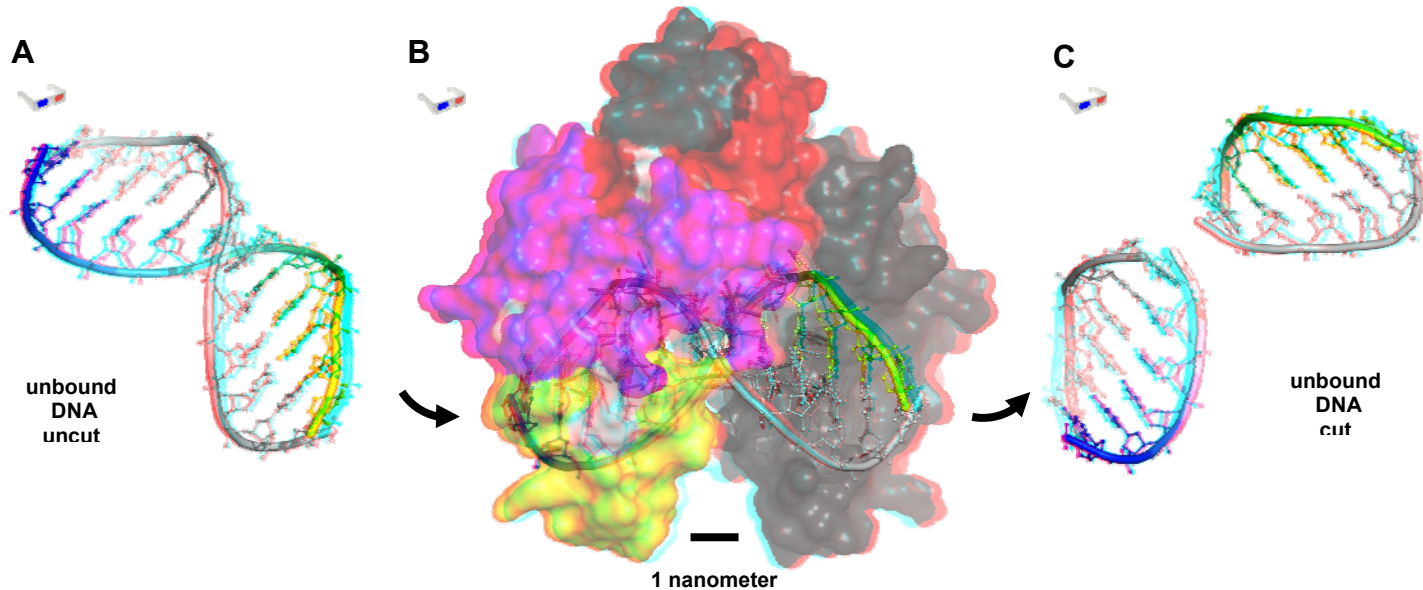
known as R.HindIII according to REase nomenclature conventions (16)) as genetic and molecular biology tools to linearize and cleave Simian Virus 40 (17-19). Together, Arber, Smith, and Nathans received the Nobel Prize in Physiology or Medicine in 1978 for the discovery of "restriction enzymes and their application to problems of molecular genetics" (20). With it, REases became the workhorses of molecular biology for manipulating and mapping specific sequences of DNA (16).

In keeping to their name, REases recognize and cleave (i.e. restrict) in between (i.e. endo-) specific short sequences, usually four to eight BPs, of DNA (i.e. nuclease) (21). They are composed of residues in nanoscale polymer peptide chains having common backbone atoms and variable side-chain atoms (Figure 2-2). The restriction sites have been studied for patterns (22) and frequency (23). These sites tend to be palindromic in sequence and the REases that cleave in between them are classified as Type IIP. The REases themselves tend to form multimers, usually homodimers, in order to accommodate the 2-fold symmetry when forming DNA-REase complexes (24) as well as other symmetries that that should be taken into computational and experimental consideration (Table 2-2). Cleavage of DNA occurs through a scissile phosphate nucleophilic attack through coordination with a metal cation cofactor (Figure 2-3). This cleavage, and the binding that precedes it, is highly sensitive to substrate sequence, cofactor, buffer and incubation

Table 2-2: Computational and experimental parallels among DNA and restriction endonucleases

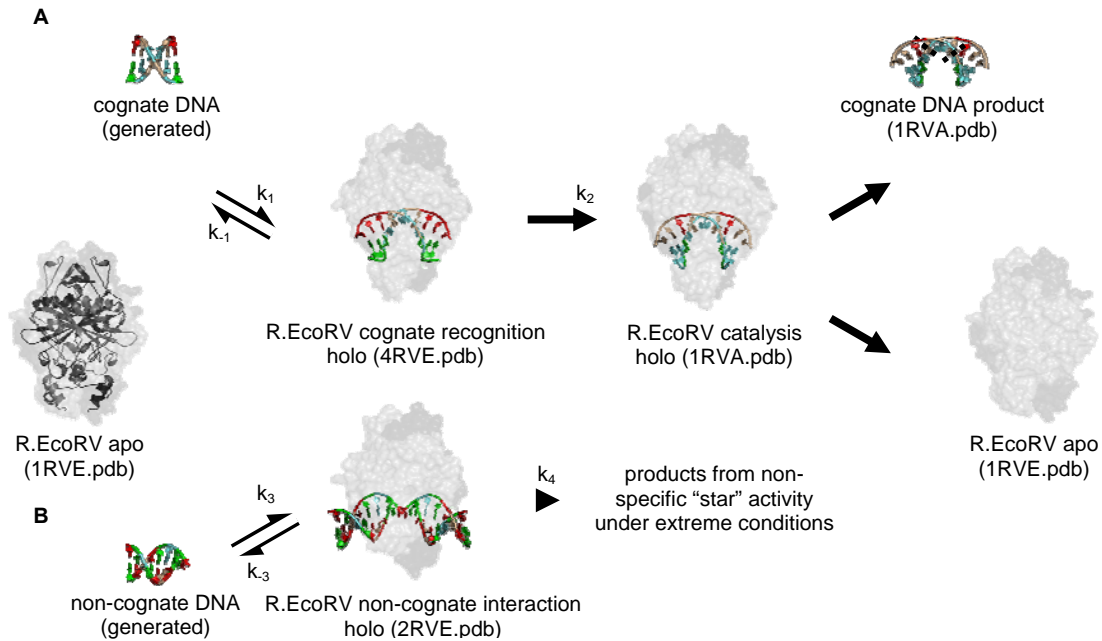
	DNA	REase
computational		
natural complex	cognate	wildtype
synthetic complex	non-cognate	mutant
structural symmetry	<i>in silico</i> structurally asymmetric palindromic anti-parallel strands	<i>in silico</i> structurally asymmetric homodimers
symmetry of mutations	engineer and evaluate both constituent strands	engineer and evaluate both constituent monomers
experimental		
natural complex	cognate	wildtype
synthetic complex	non-cognate	mutant
structural symmetry	<i>in vivo/in vitro</i> palindrome and approximately symmetric	<i>in vivo/in vitro</i> approximately symmetric homodimers
symmetry of mutations	mutate against single strand for complementary strands	mutate against single gene for both monomers

Figure 2-3: Restriction endonuclease functional activity upon cognate DNA substrate



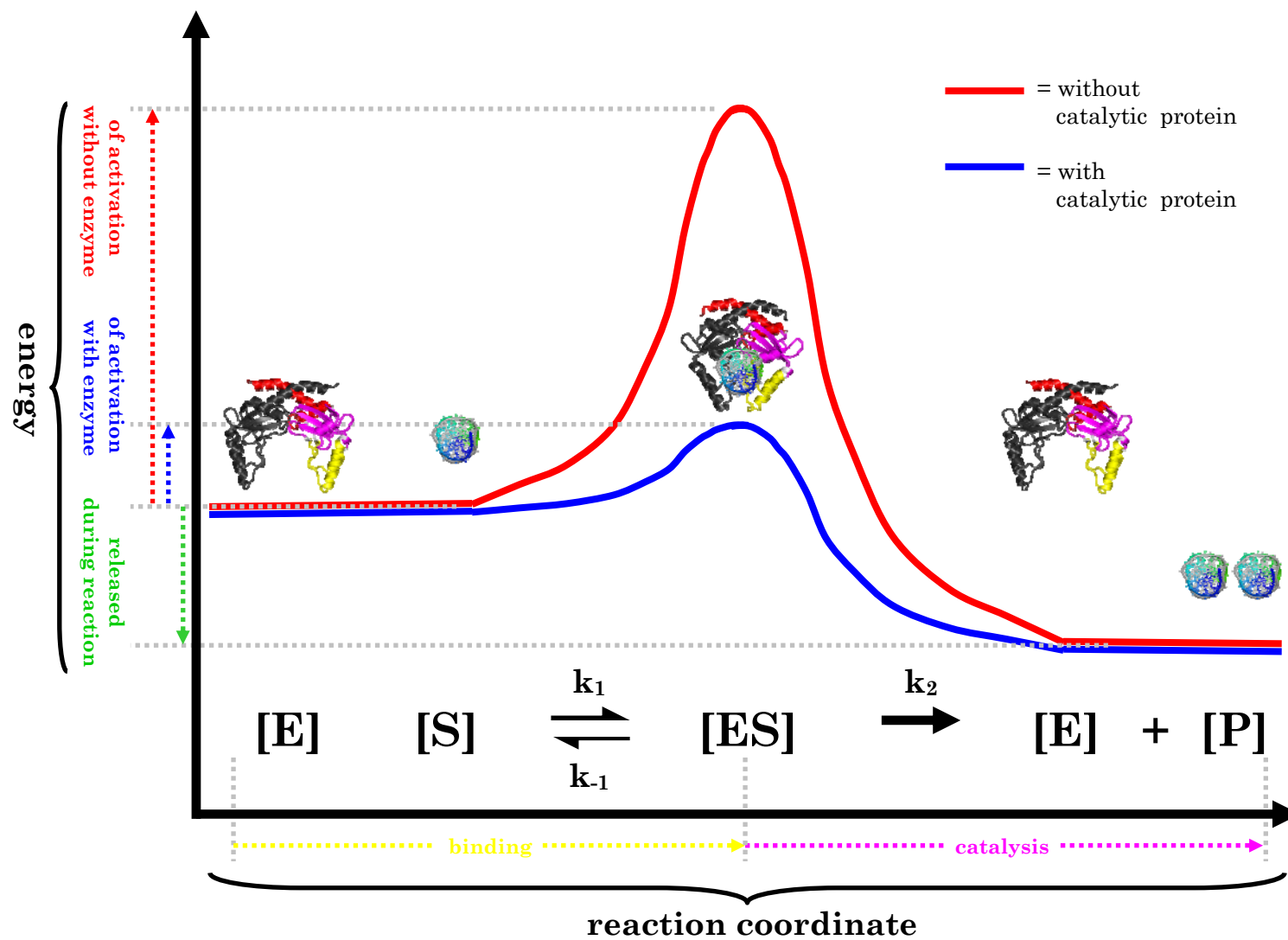
(A) DNA substrates that are uncut are probed at the nanometer scale as potential reactants through recognition directly, via the sequence of base pairs in particular DNA, and indirectly, via the sugar phosphate double helices. **(B)** The cognate DNA substrate, or one that contains a suitable base pair sequence and shape, is recognized by the REase, via conformational changes in both biomolecules that facilitate complementary recognition and coordinated catalysis of the later by the former. **(C)** DNA substrates that are cut are released non-reversibly as products of the reaction.

Figure 2-4: Restriction endonuclease binding and catalytic kinetics with cognate and non-cognate DNA substrates



In a system of reactions, spatial and kinetic aspects within DNA and between it and the restriction endonuclease (REase) are critical to recognition, deformation, and catalysis of the former by the latter. **(A)** Here, a DNA cognate sequence, 5'-gatatc-3', is complexed, k_1 , recognized, and deformed to expose the scissile phosphates for catalytic cleavage, k_2 , by RE R.EcoRV. **(B)** When non-cognate DNA is complexed, k_3 , it is not recognized, stabilized, or deformed significantly and thus is unbound, k_{-3} , or yields product release from non-specific "star" activity under extreme conditions, k_4 . The color code employed for RE homodimeric monomers is in gray shades, and for DNA adenine is in aquamarine, cytosine is in crimson, guanine is in green, and thymine is in tan.

Figure 2-5: Reaction coordinates of protein-DNA binding and catalysis



conditions (Figure 2-4); variations from normal conditions can lead to “star” activity, or non-sequence-specific cleavage of the DNA substrate. The REases serve as nanomolecular catalysts, lowering the activation energy necessary to cut DNA specifically and efficiently, but unlike many enzymes, do so without the dependence on a common biological energy currency adenosine triphosphate (ATP) (Figure 2-5).

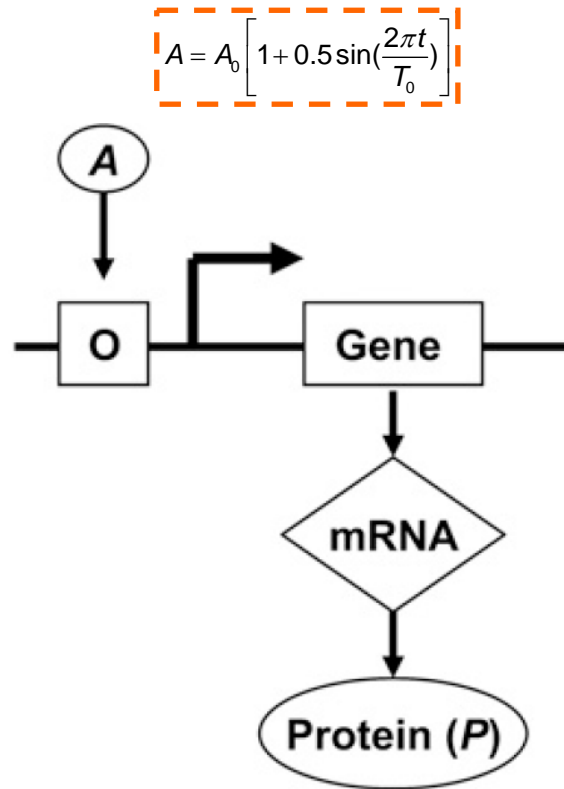
2.3. Modeling natural systems

Two distinct studies were performed on nucleic acids and the proteins that interact with them. The first discussed below involved multiple proteins and genes interacting to create a circuit, in order to model the spatial-temporal interactions among them. The second discussed below involved single proteins and genes interacting to form structural complexes, in order to model structure-function interactions for further engineering.

2.3.1. Gene-protein circuit modeling

A single stage gene circuit was modeled and signal transmission through it analyzed by mathematical modeling (25). A one-stage gene circuit was considered where an output protein (P) is driven by an oscillatory input signal (A) (Figure 2-6). In the cellular context, the input oscillations may be directly derived from environmental conditions (e.g. day-night cycles) or endogenous cellular oscillators (e.g. circadian clocks). Without loss of

Figure 2-6: Single stage gene-protein circuit model



A one-stage gene-protein circuit where a transcription activator, A , is present at certain concentrations and times. A then acts on the operator, O , of the Gene to facilitate transcription. The transcription from O creates mRNA, once again at certain concentrations at certain times. The transcribed mRNA in turn produces the output protein, Protein (P) with its own concentration-time profile. Each of these events have their own production and decay rates. In an experimental setting, Protein (P) can be a fluorescent reporter that has destabilizing mutations, such as destabilized green fluorescent protein (DsGFP) to reduce the half-time of its fluorescence and more closely indicate the off state of the transmitted signal.

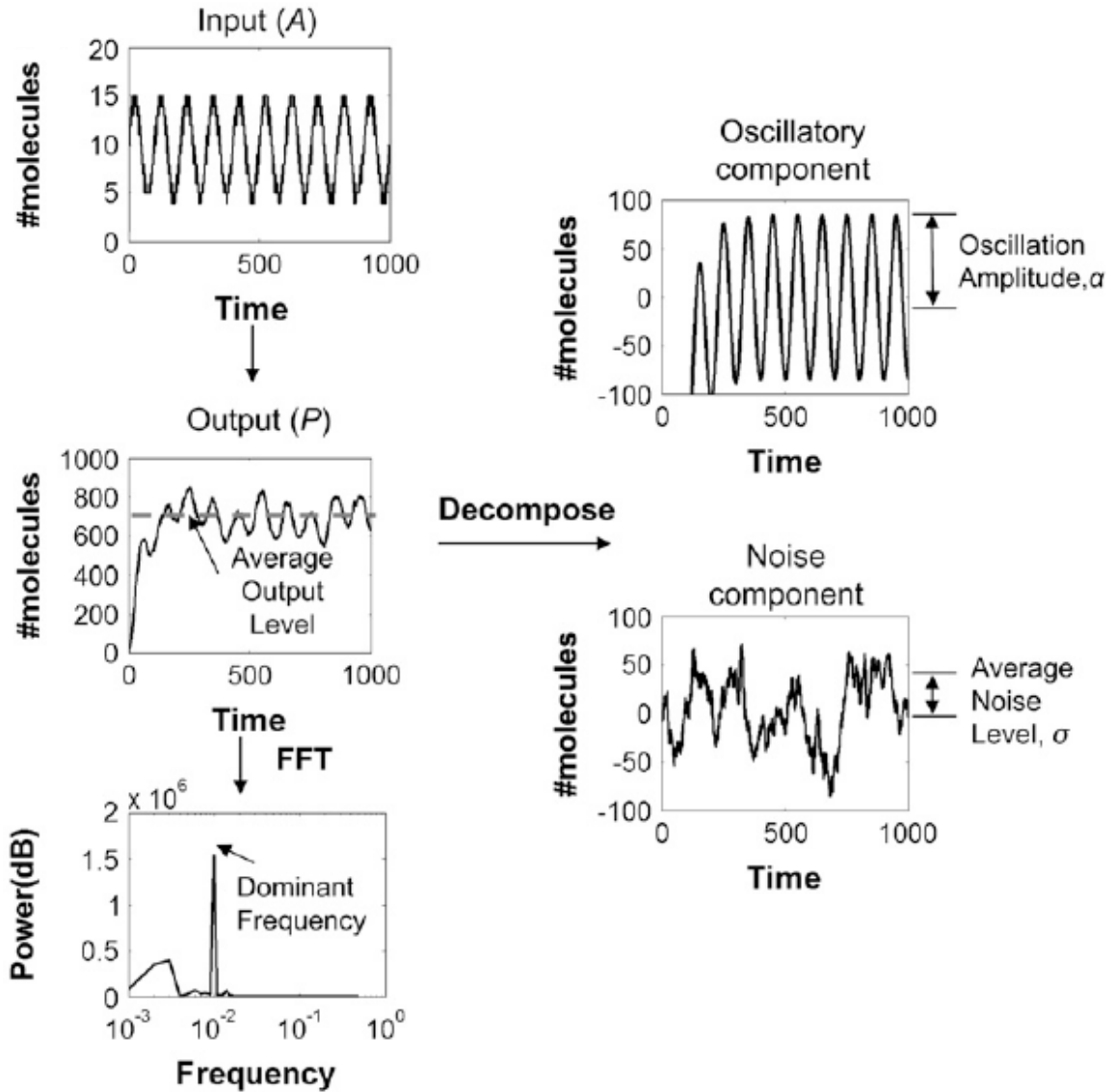
generality, it was assumed that the oscillation can be characterized by a simple sinusoid function:

$$A = A_0 \left[1 + 0.5 \sin\left(\frac{2\pi t}{T_0}\right) \right]$$

where A_0 defines both the average signal strength (A_0 is set to 10 in the example modeled herein) and the corresponding oscillation amplitude, and T_0 is the oscillation period. Two complementary analytical and numerical approaches were taken to analyze transmission of the frequency signal (Figure 2-7). In the first approach the time course of output P was decomposed into its mean and standard deviation, which is an application of the linear genetic network method (26,27). The output signal P would oscillate when the gene-protein circuit was driven by the oscillatory input signal from A. The mean is defined as the oscillatory component and the standard deviation is defined as the noise component, which tends to obscure or mask the oscillatory component. The model was hypothesized to reveal that a frequency signal is transmitted accurately if the oscillation amplitude, α , exceeds the noise level, σ . For simplicity in terminology, α and σ are termed the amplitude and noise level, respectively, of the output signal.

To complement the analytical method, numerical methods were used to analyze the P time course for its dominant frequency. If the signal transmission was accurate, this dominant frequency would be the same

Figure 2-7: An oscillatory input signal can generate an output signal with oscillations compounded with noise



The mean and standard deviation of the output signal of the linearized model was analytically computed. Here, the mean value was defined as the oscillatory component and the standard deviation as the noise component. Alternatively, the stochastic simulations of the output signal for the nonlinear system was analyzed by the fast Fourier transform (FFT) method to obtain its dominant frequency.

(within numerical errors and machine precision) as the input frequency. The dominant frequency of the P time course was calculated by using the fast Fourier transform (FFT) method. The steady-state portion of the P time course for each simulation was analyzed using the FFT method. Results from the FFT analysis were then used to extract the dominant output frequency. This output frequency would correspond to the signal frequency “perceived” by downstream processes.

By linearizing the mathematical model of the gene-protein circuit and then decomposing the output using established methods (26-28), the average output level, b , is obtained:

$$b = \frac{k_m k_p A_0}{g_m g_p}$$

where k_m is the transcription rate constant, k_p is the translation rate constant, g_m is the mRNA decay rate constant, and g_p is the protein decay rate constant. When changing a circuit parameter, the average output level was maintained at a constant 500 molecules by adjusting the k_p . For instance, if g_m is increased 10-fold, b can be kept constant by increasing k_p 10-fold. By doing so, different circuit configurations or parameter settings would on average elicit the same average level of downstream gene expression (whether or not the input frequency was maintained through transmission). The amplitude of the output oscillations, a , follows the equation:

$$\alpha = \frac{k_m A_0 k_p \sqrt{(g_m g_p - w^2)^2 + w^2 (g_m + g_p)^2}}{2(g_m^2 + w^2)(g_p^2 + w^2)}$$

where w is a function of the frequency of input signal, f_{in} :

$$w = 2\pi f_{in}$$

The corresponding average noise level, σ , equation is:

$$\sigma = \sqrt{\frac{k_m A_0 k_p (g_m + g_p + k_p)}{g_m g_p (g_m + g_p)}}$$

The amplitude of the output oscillations is defined as a decreasing function of α with increasing input frequency (f_{in}). This dependency reflects the low-pass filter characteristic of linear gene-protein circuits (29). In contrast, the σ equation is independent of f_{in} . Therefore, α would decrease below σ for sufficiently high f_{in} (Figure 2-8). In this region, frequency signals will be masked by the noise. The intersection between the σ curve and the α curve thus defines a critical frequency, f_c , beyond which the circuit will fail to transmit the input signals. For the given circuit configuration, f_c was approximately 0.02/min.

The results of the decomposition method were consistent with those from stochastic simulations. Specifically, f_{in} was varied from 0.002/min to 0.033/min. For each f_{in} , 200 stochastic simulations were performed using the Gillespie algorithm (30). The dominant frequency, f_{out} , was determined for each output time course using FFT. A parity plot between f_{in} (x-axis) and

corresponding f_{out} (y-axis) was created, and the estimated f_c (0.02/min) using the decomposition method corresponded to a transition region in the parity plot (Figure 2-9).

In most simulations, when f_{in} was less than 0.02/min, f_{out} was equal to the corresponding f_{in} . These signals were considered accurately transmitted despite cellular noise. Beyond 0.02/min, however, the average f_{out} started to deviate from the corresponding f_{in} and the deviation increased drastically with further increase of f_{in} (Fig. 2-9, striped area).

The drastic deviation was due to the fact that most output time courses gave incorrect f_{out} . The percentage of the outputs that oscillated at the correct f_{out} for each f_{in} was analyzed. This analysis quantified the fraction of a cell population that could correctly transmit the frequency signal, where behavior of each cell was represented by one stochastic simulation. It provided a quantitative measure of signal transmission fidelity for each f_{in} (Fig. 2-10). Again, the estimated f_c defined a transition point that corresponds to a drastic reduction of cells that generated the correct f_{out} . When f_{in} was less than 0.02/min, nearly 100% of the cells produced the correct f_{out} , indicating high fidelity in signal transmission. However, when f_{in} was greater than f_c , the percentage decreased drastically, indicating that the majority of cells failed to transmit the frequency signal accurately.

Figure 2-8: From analytical decomposition, the amplitude of output oscillations decreased with f_{in} , while the critical frequency, f_c , was calculated as the intersection between the “average noise level” curve and the “oscillation amplitude” curve

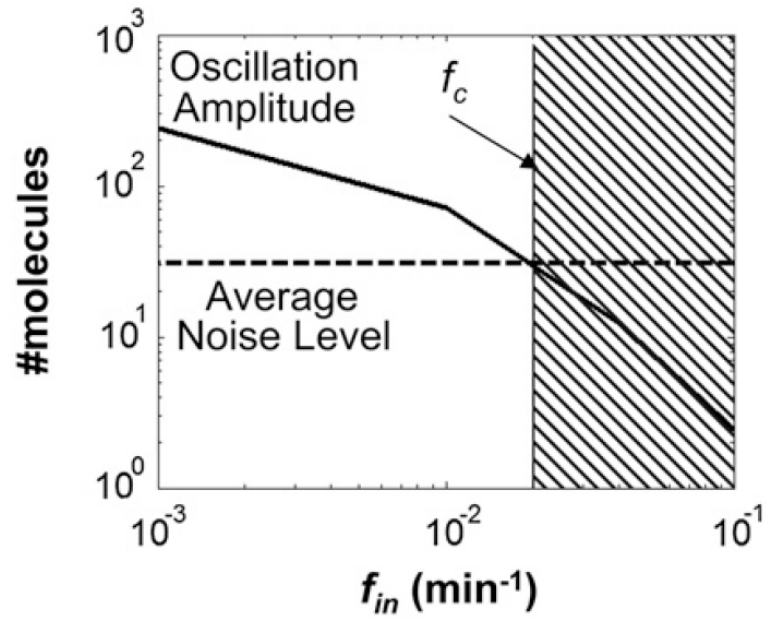


Figure 2-9: From numerical stochastic Gillespie simulation and FFT for dominant frequency, a critical frequency, f_c , consistent with the analytical decomposition was observed

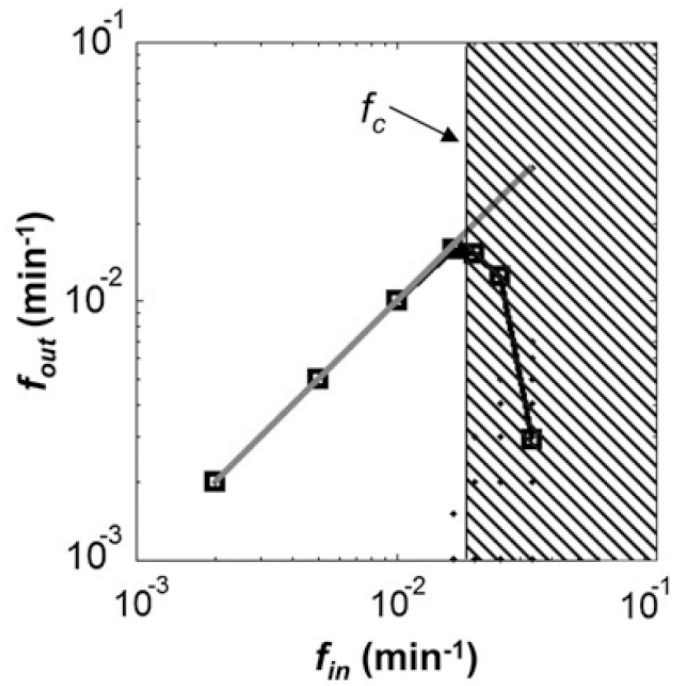
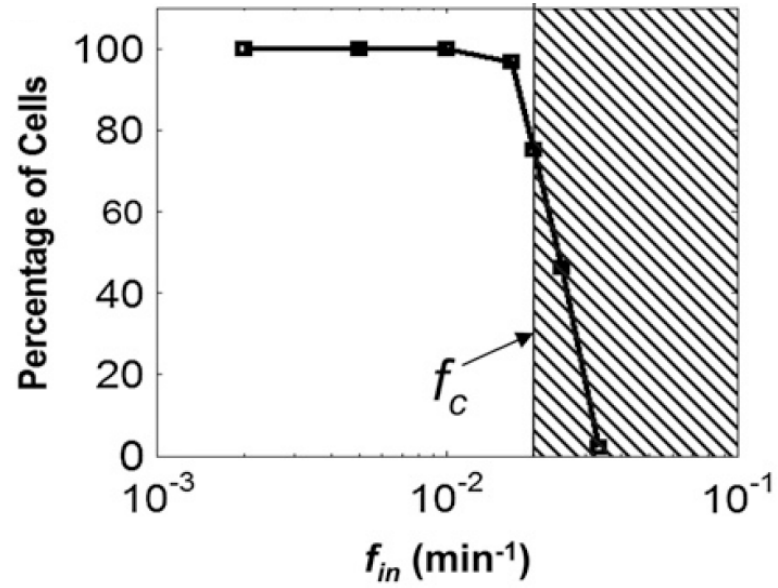


Figure 2-10: The percentage of cells that generated f_{out} with fidelity, i.e. equal to f_{in} , reduced sharply at and beyond the critical frequency, f_c



Therefore, both the analytical method and the 'brute-force method by stochastic simulation revealed an intrinsic property of frequency-signal transmission in the simple gene circuit: it is 'all-or-none', with the transmission fidelity limited by f_c . The analytical method also suggested how the f_c emerged as the interplay between the amplitude and the noise level of each output oscillation.

2.3.2. DNA-enzyme structure modeling

All uniquely available REase bound to cognate DNA structures and associated information were retrieved from the Protein Data Bank (PDB) (31,32), the Nucleic Acid Database (NDB) (33), and REBASE (34,35) for modeling (Table 2-2). Among the structures available, R.PvuII was the smallest in monomer and overall length, was available in both bound to cognate DNA and unbound forms, and with (Mg^{2+}) and without (Ca^{2+}) a catalytically competent metal cation, and thus selected as the model system for much of this dissertation (Table 2-3). For each structure, the modeling was performed to visually demonstrate the curvature of DNA and degree of distortion imparted due to the interactions with the bound REase. Since the BPs themselves are asymmetric and lack a third common point with which to establish common planes, their the vector origins O_x , O_y , O_z were determined

along three dimensions following the definitions established in (36,37). The fixed BP origin, \mathbf{F} , of the entire BP was thus defined as:

$$\mathbf{F} = (O_x, O_y, O_z)$$

All unique arrangements of single BPs were generated from ideal bond geometries using the nucgen package in the Assisted Model Building and Energy Refinement (AMBER) software suite (38,39). The \mathbf{F} for all unique arrangements of single BPs was performed, to ensure a common plane could be established (Figure 2-11). With a common plane defined by the, now, three common points in space (C1' on each base of a BP and \mathbf{F}), a structural mutation procedure of one BP to another through coordinate identification and affine space transformation and RMSD minimization between identified coordinates was possible. This was deprecated in favor of a base-specific structural mutation procedure, to accommodate for differences in separation distance between bases (40).

The \mathbf{F} was computed for all BPs in a crystallographic structure containing bound cognate DNA substrate. For comparison, the B-DNA form of the same substrate sequence was generated using nucgen and \mathbf{F} computed. In the case of R.EcoRV, the differences in the distortion between the bound cognate DNA substrate crystallographic structure and the B-DNA form were striking (Figure 2-12). The degree of distortion was clearly visualized and

Table 2-3: Restriction endonuclease crystallographic structures

REase name	bound conformations	unbound conformations	DNA recognition site and cleavage pattern	residue × multimer count
R.BamHI	4	1	5'-g ↓ gatcc-3' 3'-cctag ↑ g-5'	213 × 2
R.BcnI	1	1	5'-cc ↓ sgg-3' 3'-ggs ↑ cc-5'	238 × 2
R.BfiI	0	1	5'-actgggnnnnn ↓ -3' 3'-tgaccnnnnn ↑ n-5'	358 × 2
R.BglI	1	0	5'-gccnnnn ↓ nggc-3' 3'-cggg ↑ nnnnccg-5'	299 × 2
R.BglII	2	1	5'-a ↓ gatct-3' 3'-tctag ↑ a-5'	223 × 2
R.Bse634I	0	1	5'-r ↓ ccggy-3' 3'-yggcc ↑ r-5'	293 × 2
R.BsoBI	1	0	5'-c ↓ ycgrg-3' 3'-grgcy ↑ c-5'	332 × 2
R.BstYI	2	1	5'-r ↓ gatcy-3' 3'-yctag ↑ r-5'	203 × 2
R.Cfr10I	0	1	5'-r ↓ ccggy-3' 3'-yggcc ↑ r-5'	285 × 2
R.Ecl18kI	2	0	5'- ↓ ccngg-3' 3'-gggcc ↑ -5'	305 × 2
R.EcoO109I	1	1	5'-rg ↓ gnccy-3' 3'-yccng ↑ gr-5'	272 × 2
R.EcoRI	7	1	5'-g ↓ aattc-3' 3'-cttaa ↑ g-5'	276 × 2
R.EcoRII	0	1	5'- ↓ ccwgg-3' 3'-ggwcc ↑ -5'	404 × 2

a = adenine ; c = cytosine ; g = guanine ; t = thymine ;
n = a, c, g, or t ; r = a or g ; s = g or c ; w = a or t ; y = c or t

Table 2-2: Restriction endonuclease crystallographic structures (continued)

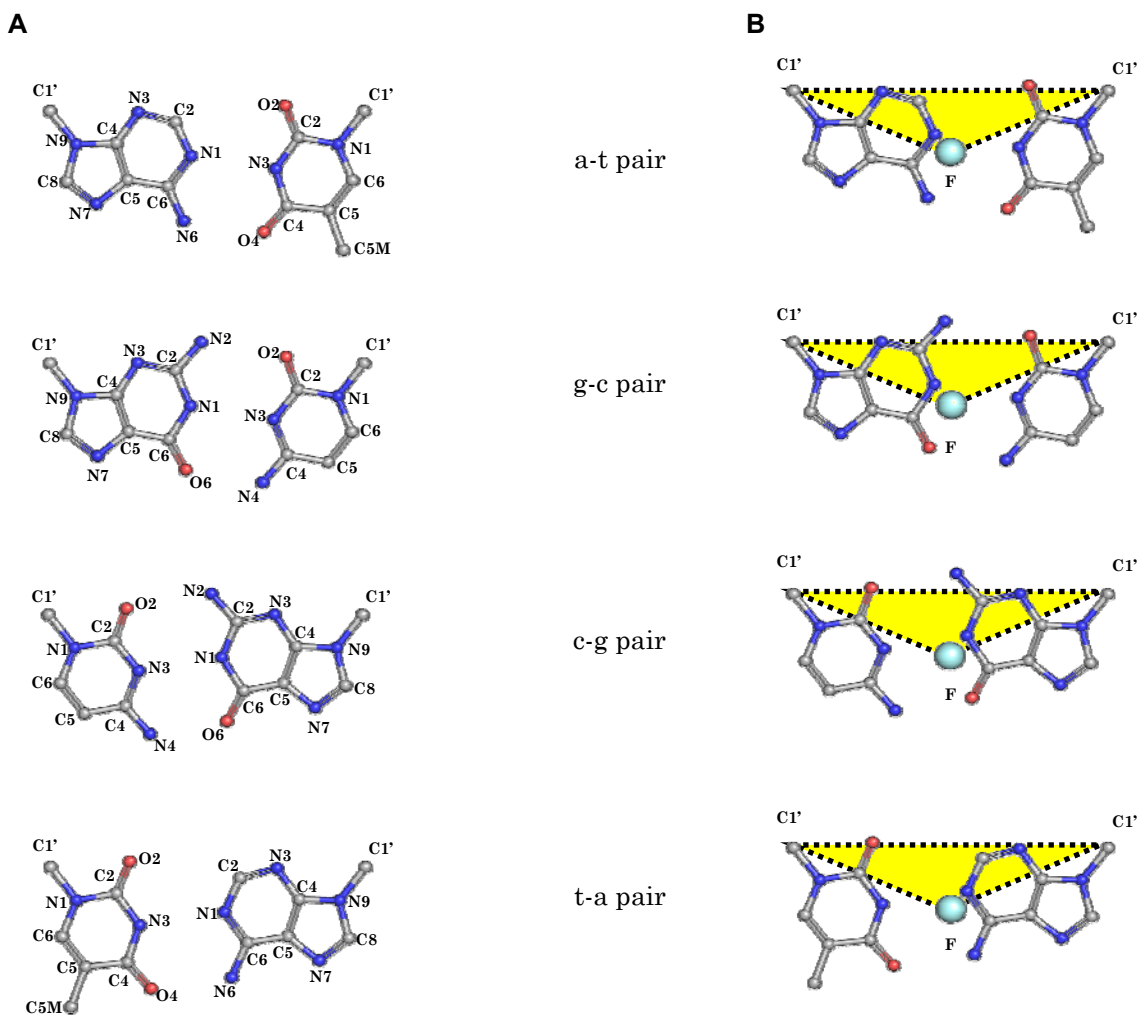
REase name	bound conformations	unbound conformations	DNA recognition site and cleavage pattern	residue × multimer count
R.EcoRV	26	3	5'-gat ↓ atc-3' 3'-cta ↑ tag-5'	244 × 2
R.FokI	1	1	5'-ggatgnnnnnnnnn ↓ nnnn-3' 3'-cctacnnnnnnnnnnn ↑ -5'	579 × 2
R.HinP1I	0	1	5'-g ↓ cgc-3' 3'-cgc ↑ g-5'	247 × 2
R.HincII	10	1	5'-gty ↓ rac-3' 3'-car ↑ ytg-5'	257 × 2
R.MspI	2	0	5'-c ↓ cgg-3' 3'-ggc ↑ c-5'	262 × 2
R.MunI	1	0	5'-c ↓ aattg-3' 3'-gttaa ↑ c-5'	202 × 2
R.MvaI	1	1	5'-cc ↓ wgg-3' 3'-ggw ↑ cc-5'	249 × 2
R.NaeI	1	1	5'-gcc ↓ ggc-3' 3'-cgg ↑ ccg-5'	317 × 2
R.NgoMIV	1	0	5'-g ↓ ccggc-3' 3'-cggcc ↑ g-5'	286 × 4
R.PabI	0	1	5'-gta ↓ c-3' 3'-c ↑ atg-5'	226 × 2
R.PvuII	4	5	5'-cag ↓ ctg-3' 3'-gtc ↑ gac-5'	157 × 2
R.SdaI	0	1	5'-cctgca ↓ gg-3' 3'-gg ↑ acgtcc-5'	323 × 2

a = adenine ; c = cytosine ; g = guanine ; t = thymine ;
n = a, c, g, or t ; r = a or g ; s = g or c ; w = a or t ; y = c or t

Table 2-4: R.PvuII crystallographic structures

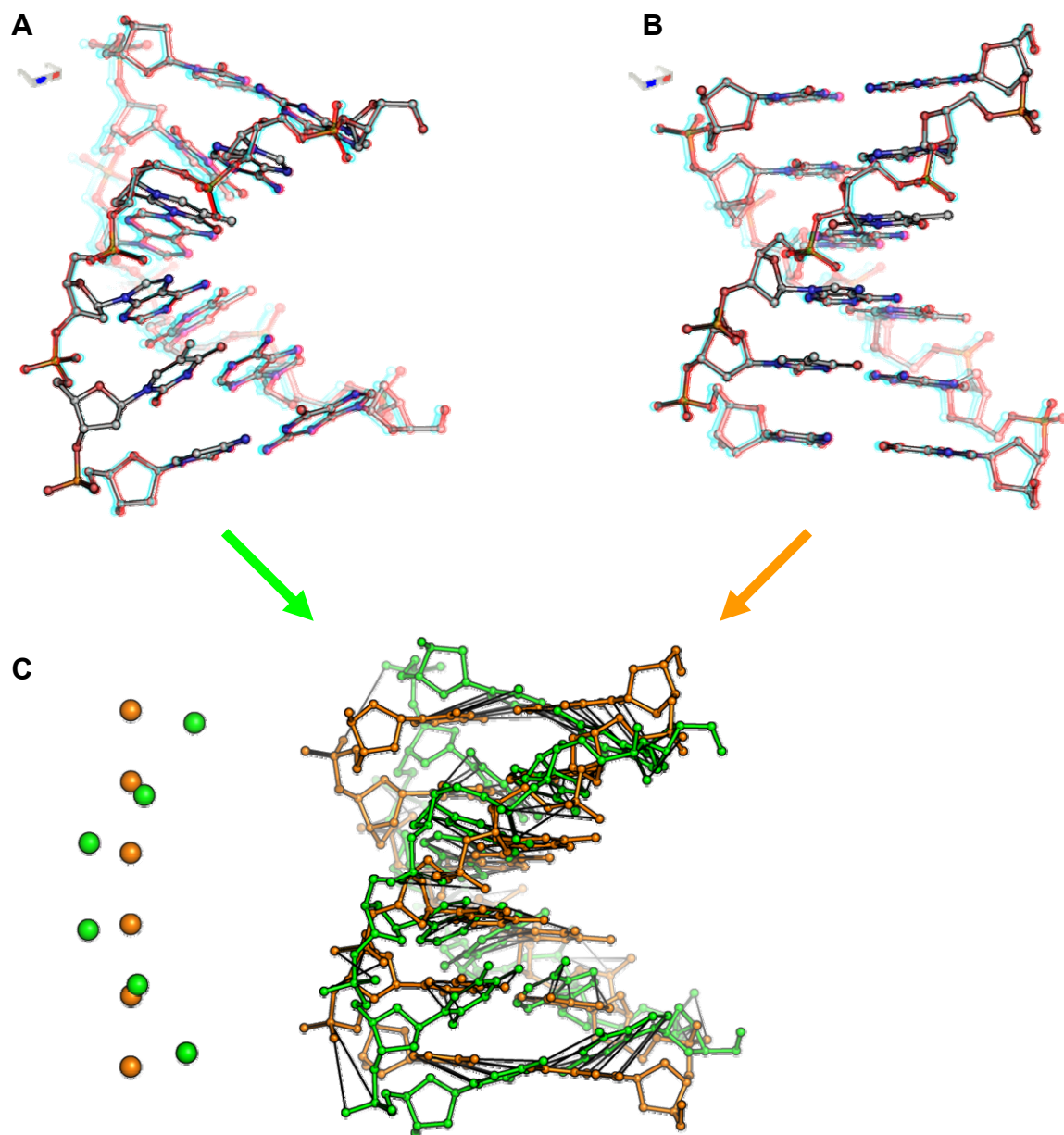
R.PvuII crystallographic structure description	resolution (Å)	PDB ID	citation
bound, with cognate DNA substrate 5'-tgaccagctggtc-3'	2.60	1PVI	(41)
bound, with cognate DNA substrate and 5-iodocytosine(4) at x 5'-tgaccagxtggtc-3'	1.76	2PVI	(42)
bound, D34G mutant with cognate DNA substrate 5'-tgaccagctggtc -3'	1.59	3PVI	(43)
bound, with Ca ²⁺ and cognate DNA substrate 5'-tgaccagctggtc -3'	1.78	1EYU	(44)
bound, with Ca ²⁺ , glutaraldehyde and cognate DNA substrate 5'-tgaccagctggtc -3'	2.50	1F0O	(44)
unbound	2.40	1PVU	(45)
unbound, Y94F mutant	2.50	1NI0	deposited in PDB but no publication
unbound, with Mg ²⁺	3.00	1H56	(46)
unbound, with Pr ³⁺ and SO ₄ ²⁻	2.05	1K0Z	deposited in PDB but no publication

Figure 2-11: Fixed basepair origin, F, for all unique arrangements of single basepairs



(A) The generated non-hydrogen heavy atoms of bases from all four B-DNA BPs, arising from AMBER force field parameters, and fiber nucleic acid studies, have two common coordinates in space, C1', but lack a third coordinate needed to define a common plane. **(B)** The coordinate identification includes computing a third common coordinate, the fixed BP origin F, occupying the 3D center of the BP and, along with C1' atoms, defining common planes among all pairs. The coordinate F was computed using X3DNA software (47-49) and visualized using the open-source PyMOL v.0.99 molecular graphics software (50).

Figure 2-12: Bound cognate DNA substrate crystallographic structure exhibits deformation compared to cognate B-DNA model



(A) Bound cognate DNA substrate crystallographic structure of R.EcoRV. **(B)** Cognate right handed B-DNA model generated using the nucgen package in the AMBER software, which is based on Arnott fiber studies (4). **(C)** Analysis of F coordinates of each of the 6 BPs in the structure (green) and model (orange) using reveal the extent of deformation when this DNA substrate is bound and spatial deviations of some corresponding atoms (black lines).

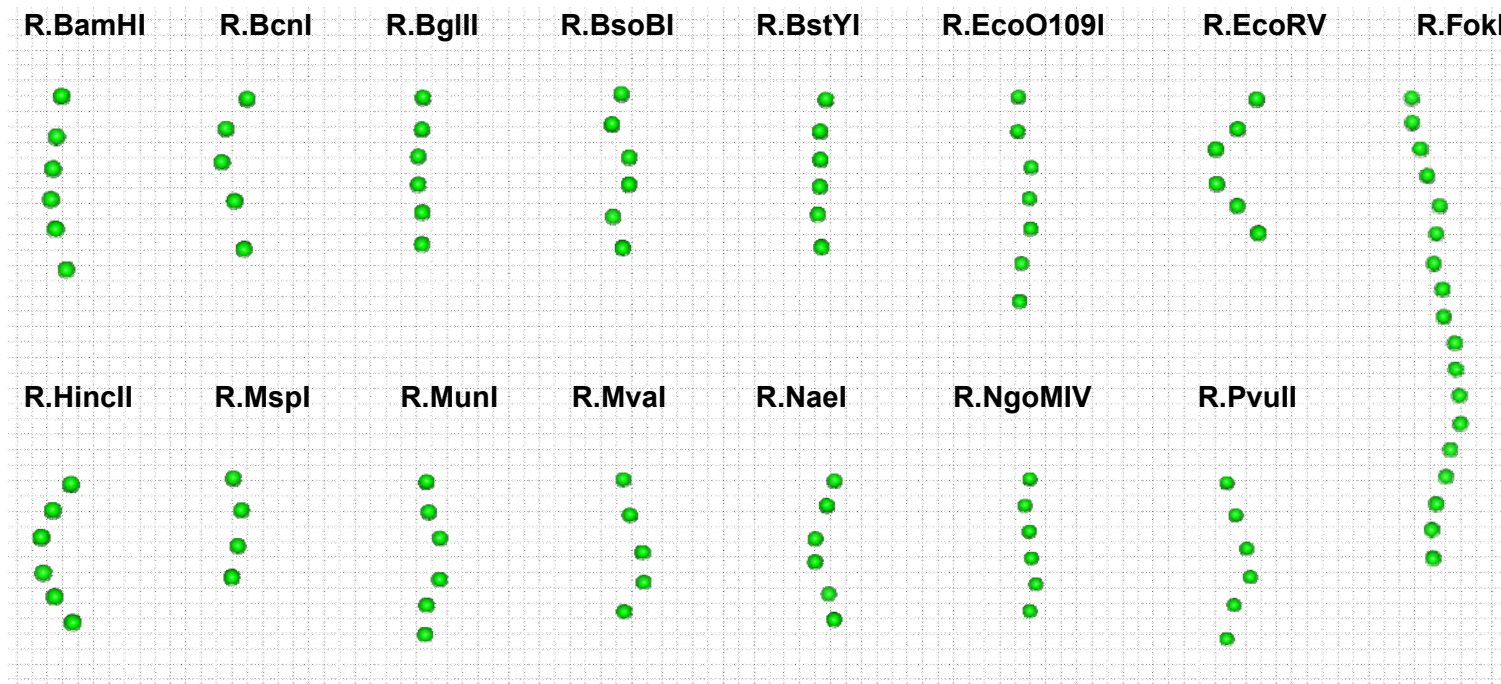
summarized by the collection of one **F** for each BP. When this analysis was repeated for many cognate DNA substrates bound to REase, it was observed that the degree of distortion differs and lacks a consistent conformation, but rarely resembled the linear arrangement of **F** in B-DNA (Figure 2-13). This modeling suggested that the overall distortion of the DNA substrate be maintained when performing structure-based molecular engineering.

2.4. Engineering of models

Engineering nucleic acid binding proteins, or other biomolecules, is non-trivial. Computational as well as experimental strategies have been employed to redesign biomolecules with desired properties.

Computational engineering of models is subject to assumptions and approximations additional to those already made in modeling natural systems. These additional assumptions and approximations are made in order to achieve the desired solutions within the available time and space of computational resources while working within the mutation tolerances of biological systems. Natural systems often adopt the lowest energy conformations possible (51). Given the input model structure, a means to model and mutate side-chains of this structure using a discrete rotamer library, and a means to measure and rank the existing and mutated structures using a pairwise energy function, determining the overall lowest

Figure 2-13: Degree of DNA substrate deformation differs among REase-bound crystallographic structures



energy conformation solution, or Global Minimum Energy Conformation (GMEC), has been proven to be NP-hard (52,53) in computational protein design (54,55). Furthermore, even approximating this solution has been shown to be NP-hard (56).

Semi-empirical and knowledge-based approaches, such as with retrospective data as support, have been applied in nucleic acid binding protein studies in order to reduce the search time and space (57). Protein-DNA interactions using structural information and evolutionary support information from databases and experiments, respectively, have demonstrated that direct interactions correlate with conservation of DNA sites, albeit with outliers (58). These interactions have also been shown to be nearly additive in energy (59). Other heuristics in conjunction with structural knowledge enabled the design of protein chimera (60) and obligate heterodimers (61). Mode analysis has also modeled thermal fluctuations and motions of proteins with some success (62-64).

Stochastic approaches have also been used for similar gains. Monte Carlo sampling methods have helped predict mutations in homing endonucleases (65,66) and of active sites (67,68). However, these and other stochastic methods, such as the aforementioned, self-consistent mean field, and genetic algorithms, are not guaranteed to find the optimal solution.

Experiment-driven search strategies have also been taken to engineering molecules for similar reasons of searching the space for functional molecules (69). Though they too do not guarantee the optimal solution, they often yield theoretical (70,71) and practical successes (72,73). For example, directed evolution and high-throughput screening methods (74,75) have found proteins having greater stability (76), improved folding (77), more thermophilicity (78) or psychrophilicity (79,80), or that are novel catalysts (81-83). Some of this directed evolution has been guided initially by computation (84,85). In altering experimental conditions around the natural REase system, some behavior modification could be achieved. In some cases, the REase was sensitive to perturbing experimental conditions in order to modify its activity (86). In other cases, a REase could be forced to comply through the incorporation of other molecules that promoted the desired activity (87,88). However, these evolution strategies require considerable experimental capacity, time, and overhead in order build and diversify representative libraries of genes, through many serial single mutations (89), and apply the proper selective pressure conditions, which are often determined through trial-and-error (90).

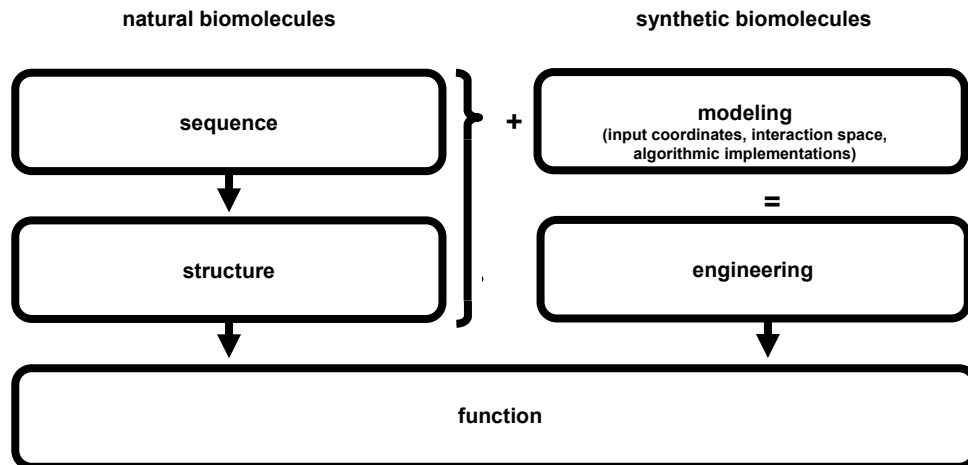
2.4.1. Operating philosophy

The operating philosophy of this dissertation is that quality modeling with state-of-the-art design algorithms can search promising spaces

efficiently to accurately predict functional engineered proteins. Our deterministic exponential-time algorithms with provable guarantees enable us to search for the GMEC and report the results with greater awareness of the search space than stochastic methods. Further developments in this area will permit systems of greater complexity to be tackled and understood. General metrics and tools, such as energy functions and rotamer libraries, enables us to extend beyond the knowledge inherent in specific systems. Here, too, improvements are ongoing and desirable. Our experimental approaches permit us to synthetically build and test the predicted mutants rapidly and as intended.

Yet, we would be remiss to not look to nature for some guidance. Cues from the central dogma of natural biomolecules are taken. The central dogma of nature states that a particular 1D sequence, whether it be a sequence of nucleic acids in DNA or RNA or of amino acids in proteins, imparts a 3D structure. It is this structure that confers some function to the DNA or protein. If the sequence and structure aspects of these nucleic acid and amino acid polymers can be modeled accurately, using suitable input coordinates, a tractable interaction space, and algorithmic implementations to perturb these coordinates in these interaction spaces, then this modeling can lead to engineering of function that may or may not already exist in nature (Figure 2-14).

Figure 2-14: Operating philosophy of synthetic biomolecules inspired by modeling and engineering central dogma of natural biomolecules



2.4.2. Input model structure visualization and preparation

For input model structures, source crystallographic structures in the form of .pdb files, which contain information such as atom types and 3D coordinates, can be retrieved from a number of sources, including the PDB (32) and NDB (33). Upon retrieval a number of cross-computing platform molecular graphics software are available that permit 3D visualization and analysis of the structure. These software include the PyMOL Molecular Graphics System (50), DeepView Swiss-PdbViewer (40), KiNG Kinemage, Next Generation (91), VMD (92), and Chimera (93).

Structures should be visualized in order to prepare and validate that input model structures are appropriate for computational engineering. Among the criteria for preparation include checking for the correct form (e.g. REase unbound or bound to DNA form), type of bound substrate (e.g. cognate, non-cognate DNA), type of cofactor and its effect on binding and catalysis (e.g. for R.PvuII, Mg²⁺ permits binding and cleavage, while Ca²⁺ permits only binding), type of chemical modifications, if any (e.g. crosslinking), and resolution. If such criteria is acceptable, each input model structure file should be further checked that only a single complex rather than multiple complexes are present, that incomplete residues are repaired to the appropriate residue or removed.

While the heavy atoms are present, hydrogen atoms can often be absent from X-ray crystallographic structures if the resolution of the structure's electron density map is too coarse to demarcate the hydrogen atom radii. When weighted to account for the frequencies of amino acids in 1,021 unrelated proteins of known sequence, there were on average 1.01 hydrogen atoms per non-hydrogen, or heavy, atom in a protein (94). Since hydrogen atoms compose about half of all atoms in a given protein, when possible, it is desirable to use a high resolution protein structure which contains hydrogen atoms or computationally supplement them using software such as Reduce (95) or WHATIF (96). With hydrogen atoms in place, further optimization to the input model structure can include flipping ring structures of side-chains asparagine, glutamine and histidine to optimize the hydrogen bonding network using software such as MolProbity (97,98). Special attention should be directed to modeling the protonation state of histidine residue imidazole moieties as they are near physiological pH (Table 2-5). The distinct protonation state should be indicated by renaming HIS residues in the input model structure to HID, HIE, or HIP, corresponding to a hydrogen atom on atom ND1, on NE2, or on both, respectively.

All protein atom nomenclature and connectivity in the prepared input structure should match the templates indicated in Appendix A.1. Amino acid

Table 2-5: Amino acid pKa values for modeling titratable moieties

amino acid	pKa
arginine	13.0
aspartic acid	4.0
cysteine	8.7
c-terminus	3.8
glutamic acid	4.4
histidine	6.3
lysine	10.4
n-terminus	8.0
tyrosine	9.6

adapted from (99)

templates, and those in the DNA should correspond to those templates indicated in Appendix A.5. Nucleic acid templates. Else, computational steps downstream will be unable to properly identify and thus perform operations upon the offending atom or residue. Caveats to the preceding statement are that the input model structure, whether representing the whole complex or a sterically reduced region, should be checked for adhering to the guidelines above to be properly identified by the downstream computations.

2.4.3. Rotamer library

The side-chains, branching from the main chain of proteins, are generally able to rotate through continuous ranges of allowable conformations. However, the computational space and time necessary to model all such conformations in a continuous range can become intractable, impractical, and possibly unnecessary. By analysis of high-resolution crystallographic structures, it was observed that there were some frequently observed conformations than others in these ranges. These observed conformations tended to be of lower energy than others in the range. The statistics of some of the distinguishing properties of these conformations, such as the distribution of side-chain dihedral angles χ_1 defined by the residues N-CA-CB-CG bonds, χ_2 defined by the residues CA-CB-CG-CD bonds, ideal covalent geometries of chemical moieties on these side-chains, and volume packing density constraints (100) were tabulated to create

rotamer libraries, that is rotated conformer libraries (101-104). In contrast, strict conformer libraries are created from geometries found in particular high-resolution crystallographic structures (105-107). The discretization of side-chain conformations in rotamer libraries allows computation to proceed in more reasonable space and time allocations, while still remaining appreciably faithful to experimental observations. Appendix A.2. Amino acid rotamer library and Appendix A.3. Amino acid rotamer volumes were used in this dissertation.

2.4.4. Energy function

The conformations adopted by side-chains and mainchain of the protein and by the DNA substrate have associated energies. To compute these energies, an empirical molecular mechanics energy function can be applied. This energy function is often composed of a sum of terms that each represent various biophysical interactions (108-110). These include dihedral bonded interactions and van der Waals steric, Coulombic electrostatic and hydrogen bonding non-bonded interactions which can vary when molecular engineering the input models, such as those found in the Assisted Model Building with Energy Refinement AMBER energy function (39). Since the energy computation occurs between pairs of all atoms in the input model, the terms of this energy function are necessarily pairwise-decomposable.

Like the rotamer library, the energy function attempts balance the sophistication necessary to capture some of the biophysical interactions of molecules reasonably faithfully and the simplicity needed for computations using the function to successfully complete in reasonable periods of time. For example, hydrogen bonding is a biophysical interaction used by sequence-specific nucleic acid binding proteins in order to directly recognize the nucleic acid substrate bases and to indirectly recognize other substrate features such as phosphates on the substrate BB (5). However, a 10-12 hydrogen bonding term has been shown to be unnecessary due to improvements in the terms that model van der Waals steric and Coulombic electrostatic interactions (39). Another important biophysical interaction is that imparted by water on other molecules. Like the continuous range of side-chain rotations, Explicit solvation, or representation of the biophysical effects of water on other molecules, can be computationally expensive to model. An implicit solvation model, such as the Lazaridis-Karplus effective energy function (EEF1) has been demonstrated to reasonably approximate the results from explicit modeling of water molecules while maintaining a lower computational cost (111). Furthermore, similar to the other aforementioned energy function terms, the EEF1 implicit solvation term is pairwise-decomposable. Appendix A.4. Energy function was used in this dissertation.

2.5. Discussion

In this dissertation, we computationally extended and applied state-of-the-art design algorithms to NABPs and enzymes in the Donald Laboratory and experimentally synthesized and evaluated the designed sequences and structures in the Tian Laboratory, both at Duke University, Durham, NC, USA. In subsequent Chapters, where possible, we have used readily available, open-source software and developed software for open-source release. Modeling software such as the aforementioned PyMOL and DeepView Swiss-PdbViewer were used for input model structure visualization and preparation, particularly in Chapter 3. Design algorithms such as minimized side-chain DEE (minDEE) (112-114), in combination with A* (115), which is a branch-and-bound algorithm for gap-free list enumeration, and K*, which is a provably-accurate ensemble-based scoring algorithm were extended to NABPs in Chapter 4. Documentation and open-source release of this software with extensions is discussed in Chapter 5.

3. Single-conformation engineering of nucleic acid binding proteins

Certainly no subject is making more progress on so many fronts than biology, and if we to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jiggling and wiggling of atoms.

– Richard P. Feynman
American physicist

This chapter has been adapted from a manuscript that was joint work with Peijun Zuo and Jingdong Tian:

Reza F., Zuo P., Tian J. Protein interfacial pocket engineering via coupled computational filtering and biological focusing criterion. *Annals of Biomedical Engineering: Special Issue: Systems Biology, Bioinformatics, and Computational Biology*. 2007, 35: 1026-1036.

and partially from a research meeting abstract that was joint work with Peijun Zuo and Jingdong Tian:

Reza F., Zuo P., Tian J. Theoretical and empirical perturbations of endonuclease-DNA biomolecular complexes. *Duke University Center for Biomolecular and Tissue Engineering Kewaunee Event*. 2007, Durham, NC.

3.1. Motivation

To engineer bio-macromolecular systems, protein-substrate interactions and their configurations need to be understood, harnessed, and utilized. Due to the inherent large numbers of combinatorial configurations and conformational complexity, methods that rely on heuristics or stochastics, such as practical computational filtering (CF) or biological focusing (BF) criteria, when used alone rarely yield insights into these ensembles or successes in (re)designing them.

Here we use a coupled CF-BF criterion upon an amenable interfacial pocket (IP) of a protein scaffold complexed with its substrate to undergo a proper set of residue replacement and R-group refinement (R^4) to filter out energetically unfavorable residues and R-group conformations, and focus in on those that are evolutionarily favorable.

We show that this coupled filtering and focusing can efficiently provide a putative engineered IP candidate and validate it computationally and empirically. The CF-BF criterion may permit holistic understanding of the nuances of existing protein IPs and their scaffolds and facilitate bioengineering efforts to alter substrate specificity. Such approach may contribute to accelerated elucidation of engineering principles of bio-macromolecular systems.

3.2. Overview

Elucidating the properties of a protein interfacial pocket (IP) can be a daunting task (52,53), let alone re-engineering it by altering residue and R-group arrangements to endow intended new functionalities. The IP of a protein may be abstracted as a set of amino acid residues, not necessarily adjacent in the linear polypeptide sequence, that form a local biochemical environment in three dimensional (3D) structure using conformations of their R-groups that is favorable to binding the proper substrate (116). These residues are housed amongst the rest of the residues, or scaffold, that do not participate in binding. Contribution to this favorable local environment can arise from a number of biophysical factors. The steric effects, approximated by a pairwise Lennard-Jones interaction energy, E^{vdW} , of the van der Waals (vdW) radii of atoms composing the amino acid residues of the IP as well as the proper substrate, contribute to favorable binding between them and provide hindrance and shielding against unintended side-reactions involving other substrates (117). The Coulombic effects, represented by an analogous pairwise electrostatic interaction energy, $E^{electrostatics}$, enable charge complementarities between regions of the IP and proper substrate and repulsive mismatches with other substrates (117). And since both the IP and a region of the substrate with which it interfaces are occupied by electrostatic interactions with water molecules or some other solvent, evacuating this

solvent is quantified as the electrostatic desolvation energy of the former, $\Delta G^{desolvation\ IP}$, and the latter, $\Delta G^{desolvation\ substrate}$. Thus, an estimate has been often used for the net binding free energy, ΔG , from a linear sum of these weighted energies (117):

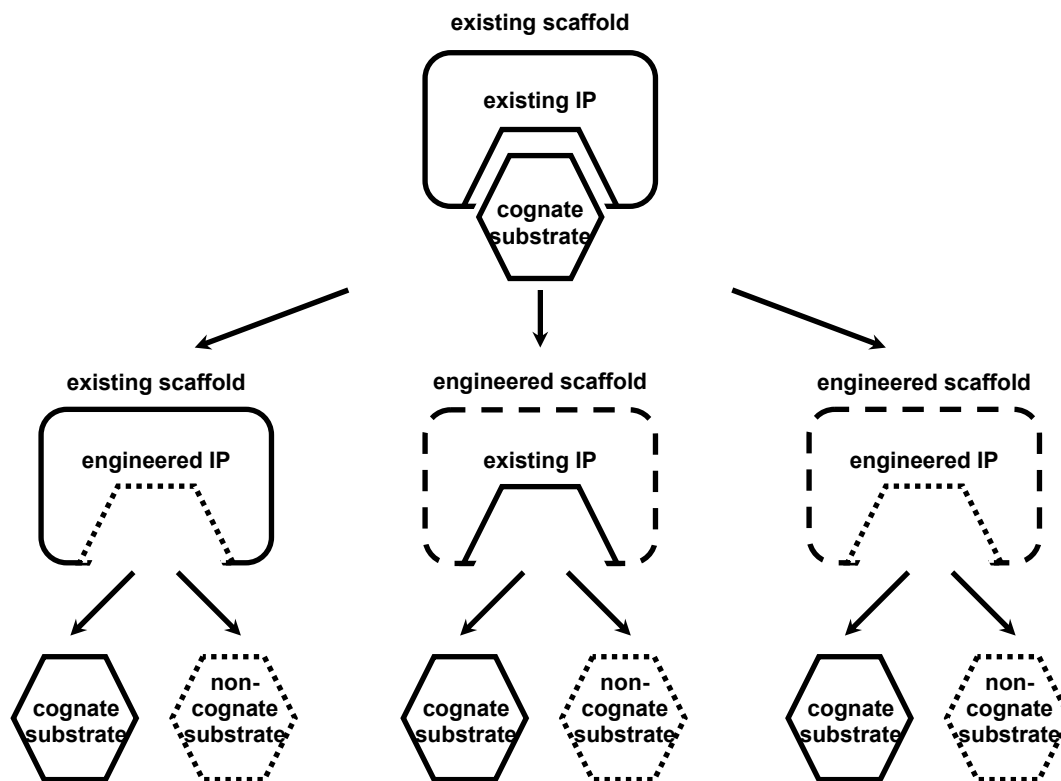
$$\Delta G = w_{IP}^{desolvation} \Delta G_{electrostatics}^{desolvation\ IP} + w_{substrate}^{desolvation} \Delta G_{electrostatics}^{desolvation\ substrate} + \sum_i w_i^{VDW} E_i^{VDW} + \sum_i w_i^{electrostatics} E_i^{electrostatics} + C$$

The protein IP engineering possibilities, as outlined in Figure 3-1, and IP residue replacement R-group refinement (R⁴) maintain these favorable ΔG of interactions and global minimum energy conformations (GMECs) of the protein-substrate complex as a whole (118).

3.2.1. Computational filtering approaches

There are a number of CF approaches to perform R⁴ with the aforementioned energetic conditions under consideration. However, due to the rapidly increasing degrees of freedom at each residue, n , of the protein chain, coupled with the specific characteristics of the 20 amino acids that can be found at each position, an colossal combinatorial quagmire of 20^n possibilities require modeling and analysis—and for an average-sized protein composed of 100 amino acids, simulating 20^{100} possible physical combinations exceeds the number of known atoms in the universe. Thus, the probability of the protein's IP locating its native state by pursuing all these combinations is

Figure 3-1: Protein interfacial pocket engineering possibilities



The existing (possibly wildtype) protein, consisting of the existing scaffold and IP with the cognate substrate (top) can serve as a starting point for three distinct engineering possibilities: an advantageous existing scaffold can support an engineered IP that binds the cognate or non-cognate substrate differently than wild-type (bottom left); an existing IP that binds the cognate or non-cognate substrate well can be adapted to an engineered scaffold (bottom center); or both IP and scaffold can be engineered to provide advantageous support and binding to a cognate or non-cognate substrate (bottom right).

biologically infeasible (known as the Levinthal paradox) (119) and computationally impractical (known as the Blind Watchmaker paradox) (119). An exhaustive structural bioinformatics search for IP formation and end-state continues to be a challenge that is tackled using filtering, heuristics, homology, distributed computing, and high performance supercomputers with varied success (120).

Heuristics are often helpful and necessary in undertaking R^4 at the scale of IPs. For example, heuristics in genetic algorithms, mean field algorithms, constraint logic programming enumeration, or database search perform adequately under certain scenarios and assumptions and not as well with others (121). While the computational cost is lessened or efficiency increased compared to the exhaustive search, the quality, however, of the end solution may or may not be consistent rather than assuring that the particular IP R^4 generated by the heuristic is located at the GMEC.

Homology can often aid in proper R^4 as well. Here, informatics searches and interpolations from signature sequences of a few residues composing a key motif of IP, substrate, or both can provide clues for engineering. This can extend further to domain sampling of entire regions across the protein that compose the IP. While this may be effective in well-investigated and documented systems, those sequences or structures with no similarity or availability of such information can hinder this approach. Even

with fertile sources, often the R^4 is limited by what has been already observed to transpose well (122,123).

In similar fashion, partitioning and docking can narrow the possibilities for R^4 (124). A collection of IP conformations can be generated that each present a different vdW, electrostatic profile or desolvation cost. By docking this collection of IP conformations to the proper substrate, the affinity features of those subpartition of IPs that dock more readily can be gleaned. However, fully enumerating all the elements in this collection may be computationally difficult or biologically unsubstantiated.

Furthermore, exact filtering algorithms, among them integer programming (125), dead-end elimination (DEE) (126-129) (130-132), and A* (115), advances the R^4 process by eliminating and enumerating possibilities (133). Dead-End Elimination (DEE) is a provably-accurate deterministic algorithm guaranteed to find the optimal solution, if such a solution exists (126,128,134). Like many of the aforementioned approaches, DEE reduces the search space by pruning rotamers, placed using the aforementioned rotamer library, that are provably cannot be part of the global minimum energy conformation (GMEC). This pruning is accomplished by using the aforementioned energy function to calculate the rotameric energy of interactions between rotamers and the backbone, between rotamers and itself, and between rotamers and other rotamers. In DEE the relative global

energy, E_{global} , of an IP is composed of the linear sum of the energy contributions from the backbone, the self and interaction with backbone energy, $E(i_r)$, of rotamer, r , at its position, i , and its pairwise interaction energy with rotamer at nearby position, j :

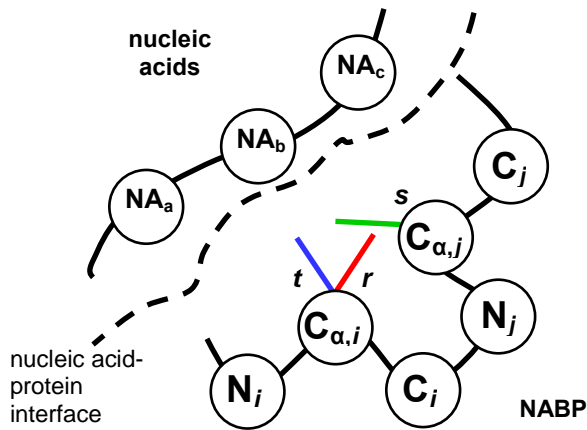
$$E_{global} = E_{backbone} + \sum_i E(i_r) + \sum_i \sum_j E(i_r j_s); \quad i < j$$

Thus, if the minimum energy, determined by via some given discrete rotamer library and energy function, or best case, arrangement of rotamer, i_r , still has a higher energy than the maximum energy, or worst case, arrangement of an alternate rotamer, i_t :

$$E(i_r) + \sum_j \min_s E(i_r j_s) > E(i_t) + \sum_j \max_s E(i_t j_s); \quad i \neq j$$

then the former rotamer is considered an energetic dead-end for further investigation as it and its variant arrangements are guaranteed to not be a participant in the GMEC, thus filtering the number of possibilities than need to undergo R^4 (Supplementary Figure 3-1). The conformations can be enumerated in a gap-free list, ordered by lower-bounds of energies, using computational techniques such as A* branch-and-bound search. DEE was originally applied by Desmet, Maeyer, Hazes and Lasters for side-chain replacement and refinement (127). Mayo and co-workers adapted the DEE for sequence search and biologically focused on different protein architectures

Supplementary Figure 3-1: Traditional dead-end elimination (DEE) criterion for nucleic acid binding proteins



$E_{backbone}$ = backbone energy
 $E_{-}(i_r)$ = rotamer-backbone, intrarotamer energy
 $E_{-}(i_r, j_s)_{i < j}$ = pairwise rotamer energy

traditional DEE criterion

$$E(i_r) + \sum_j \min_s E(i_r, j_s) > E(i_t) + \sum_j \max_s E(i_t, j_s) \quad i \neq j$$

for redesign, such as core, boundary, or surface (135). The computational guarantees provided by DEE and A* can be accompanied by significant computational cost if DEE pruning is ineffectual and enumeration of all elements from the rotamer library in use at each residue position of the IP must be done. In addition, since these IP R-groups need to be energy minimized as a whole, then DEE may no longer be provably-accurate. In this Chapter, we will biologically focus but on different protein levels, and then proceed to extending and applying a more powerful minimized DEE criteria (minDEE) with side-chain flexibility that is provably accurate for nucleic acid binding proteins in Chapter 4 (112-114). In summary, the CF approaches are often a trade-off between the quality of the end IP candidate and the efficiency to reach it (84).

3.2.2. Biological focusing approaches

Correspondingly, there are many BF approaches to perform R^4 so that resulting possibilities are in or near the aforementioned energetic conditions, perhaps by virtue of the constraints and fitness requirements existing in and imposed by the biological environment (69,85). Here, the parallel processing nature of this environment may provide a natural, even advantageous, platform to evaluate the large combinatorial number of possibilities and interdependencies to be considered in a tractable manner. However, this evaluation is often performed in a stochastic, discovery-driven investigation

using various mutagenesis techniques, recombination, and directed evolution among others to screen for high performing clones or select those that survive from a large starting population representing the number of possibilities (136).

Stochastic methods are often necessary for R^4 at a single position in the protein, let alone the half-dozen to a dozen residues that comprise some IPs (137). Consider a random mutagenesis methodology using mutagenic chemicals, wobble base PCR, or error prone PCR to incorporate mutations at the genetic level that will be selected or screened for the desired characteristics at the protein level. Though apparently misguided, it has been observed that non-obvious mutations can give rise to proteins with new characteristics (138).

Another set of approaches to achieve R^4 using biology relies on using the recombination of existing components in the system to generate new promising possibilities (139). Among these is incremental truncation to correlate the loss or gain of certain IP features and functions to the gene and protein truncation positions (140,141). There is also homologous gene shuffling to generate variants of the original IP from internal wellsprings of diversity (142).

These external and internal sources of stochastics can be considered aspects of directed and simulated evolution, which mimic the fitness

requirements, survival and natural selection, propagation and amplification and of individuals, or IPs, to evaluate massive potential-filled populations with desirable properties (137). However, these stochastic approaches rely on the robustness of this evolutionary condition to propagate order from randomness. In summary, the BF approaches are usually a compromise between the intended end IP and those that arise serendipitously or survive having unintended properties.

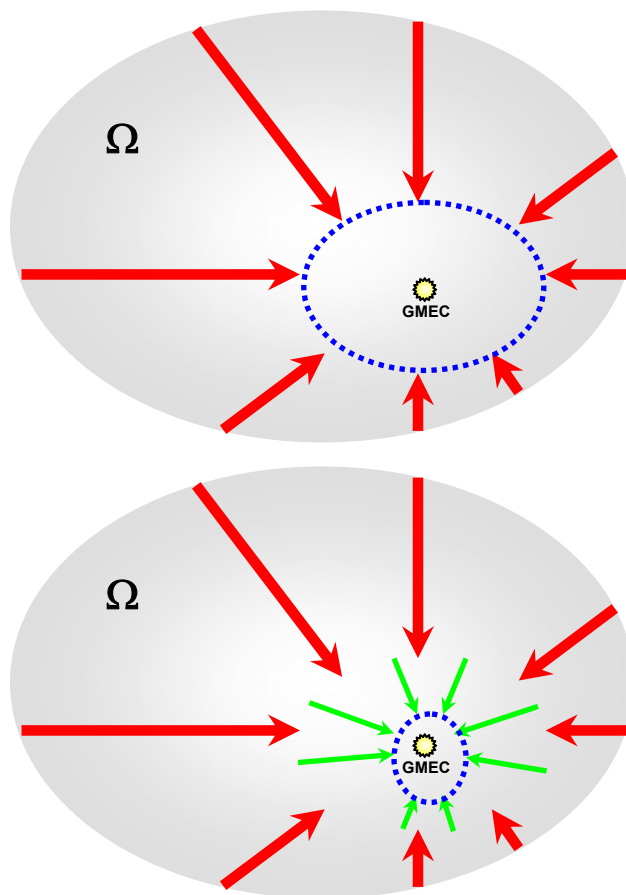
3.2.3. Coupling computational and biological approaches: CF-BF

While CF and BF each has its own advantages and drawbacks, a synergistic coupling of CF and BF may narrow the scope to a smaller number of high-quality, intended candidates more efficiently than either alone (Figure 3-2). This smaller number of possibilities is also more amenable to downstream computational and empirical evaluation and feedback. In this research, we demonstrate the application of the CF-BF criterion to computationally engineer a putative IP on the scaffold of the restriction endonuclease R.PvuII to bind the DNA substrate of a different restriction endonuclease R.EcoRV.

3.3. Methods

3.3.1. Molecular system selection










Figure 3-2: CF-BF reduces the search space and the corresponding cost required to locate the global minimum energy conformation



Using existing computational filtering criterion, shown in red arrows, the search space of all possibilities, Ω , eliminates residues and R-group configurations of those residues that are most likely not in the GMEC based on pairwise local energies, to yield a smaller number of conformational possibilities, shown in dotted blue line, that must be evaluated via global energy minimization (top panel). Coupling to a biological focusing criterion, shown in green, improves this condition by further reducing Ω to an even smaller number of possibilities based on evolutionarily relevant residues and R-groups, to be evaluated for minimum global energy as well as functionality (bottom panel).

As the various CF approaches are described in aforementioned references, the materials and methods for the BF aspects of the CF-BF criterion are as follows. Since restriction endonucleases are representative class of proteins with high specificity to their DNA substrates (143,144), the electronic .pdb file containing crystallographic coordinates of protein structures of comparative candidates were obtained via a survey of the Protein Data Bank (<http://www.rcsb.org/pdb>) (32) and REBASE: the Restriction Enzyme Database (<http://rebase.neb.com>) (34) online resources. Of the 22 restriction enzymes and 10 methyltransferases with available crystal structural information, PvuII was the only candidate for which structural information in .pdb files were available for both the restriction endonuclease, R.PvuII, (denoted by the R. prefix) and the corresponding methyltransferase modification enzyme, M.PvuII, (denoted by the M. prefix), and thus were downloaded for its greater comparative specificity of the restriction-modification system for the same DNA substrate (145,146). Also, as seen in Table 3-1, given that R.PvuII is among the smallest known restriction endonucleases available, it may be more amenable to structural bioinformatics analysis and a tolerant acceptor of IP engineering (41). The other candidate .pdb files downloaded were R.EcoRI (147), R.EcoRV

Table 3-1: Comparative protein candidate structures

protein	description	PDB ID	structure	multimers and residue count
R.PvuII- N	PvuII N ative unbound form restriction enzyme	1PVU		dimer (shown) with 157 residues per monomer
R.PvuII- D	PvuII restriction enzyme with D NNA substrate	1PVI		as above with DNA base pairs
M.PvuII- N	PvuII cytosine N-4 apo form M ethyltransferase	1B00		monomer (shown) with 323 residues
R.EcoRI- N	EcoRI N ative apo form restriction enzyme	1QC9		dimer with 276 residues per monomer (shown)
R.EcoRI- D	EcoRI restriction enzyme with D NNA substrate	1ERI		as above with DNA base pairs
R.EcoRV- N	EcoRV N ative apo form restriction enzyme	1RVE		dimer (shown) with 245 residues per monomer
R.EcoRV- D	EcoRV restriction enzyme with D NNA substrate	4RVE		as above with DNA base pairs
R.BamHI- N	BamHI N ative apo form restriction enzyme	1BAM		dimer (shown) with 213 residues per monomer
R.BamHI- D	BamHI restriction enzyme with D NNA substrate	1BHM		as above with DNA base pairs

(148-150), and R.BamHI (151) since these restriction endonucleases also have available structural information. For all these candidates, except M.PvuII, both the unbound, or native, (denoted by the -N suffix) and the bound (to DNA substrate, denoted by the -D suffix) forms were accessible, providing further comparison of conformational change with binding and a fertile source of donor IP residues for engineering. While it would have been insightful to compare isoschizomers, or enzymes that recognize the same DNA substrate and performs similarly, to R.PvuII's 5'-cagctg-3' such as R.DmaI, there are no such available crystal structure data available at this time. Similarly lacking crystal structure data but even more useful from a comparative perspective would be neoschizomers, or enzymes that recognize the same DNA substrate but perform their activity at different positions from the prototype (152).

3.3.2. Primary structure BF

Primary structure (PS) BF analysis on the proteins (PDB IDs: 1PVI, 1BOO, 1ERI, 4RVE, 1BHM) was performed by first querying the REBASE database for these comparative protein candidates and extracting the source organism. For each protein and associated source organism, the identity of the corresponding oligonucleotide bases of DNA substrate and means of interaction was determined. The one dimensional (1D) protein polypeptide sequences were retrieved from the Protein Data Bank for each .pdb entry and

cross-checked with the SEQRES fields in the .pdb files. Furthermore, these sequences were used to construct a phylogenetic tree. Multiple sequence alignments of the sequences were generated using a pairwise alignment evolutionary distance matrix, neighbor-join clustering and CLUSTALW algorithms (153).

3.3.3. Secondary structure BF

Secondary structure (SS) BF analysis was performed for the remaining proteins (PDB IDs: 1PVI, 1ERI, 4RVE, 1BHM) after PS BF by querying the Protein Data Bank for “Sequence Details” section to assign secondary structure based on the .pdb structure file’s “Author” and domain assignment using the Structural Classification of Proteins (SCOP) backend database (154). The hydrophobic profiles of each protein were determined using the Kyte-Doolittle method (155) .

3.3.4. Tertiary and quaternary structure BF

Tertiary (TS) and quaternary (QS) structure BF analyses was performed for the remaining proteins (PDB IDs: 1PVI and 4RVE) after PS and SS BF using the open-source PyMOL v.0.99 molecular graphics real-time visualization and manipulation software with embedded Python scripting and interpreter (50). For each PDB ID, the corresponding .pdb coordinate file was loaded, preset to cartoon rendering of the polypeptide SS, TS, and QS, enabled main and side chain rendering of the oligonucleotide substrate, and

directional coloring of the polypeptide chains using a spectral gamut ranging from cooler blue hues at the N-termini to warmer red hues at the C-termini. After isolating all atoms composing the 6-mer recognition sequence on the sense oligonucleotide chain to serve as points of origin, the set all residues within a 3.0 angstroms (\AA) boundary from these points were selected. This set was pruned of those atoms located at the origin and the anti-sense oligonucleotide chain, leaving the subset of atoms that were part of IP residues and R-groups within this boundary. Upon labeling, the polypeptide positions and residues at those positions were tabulated against the closest proximity oligonucleotide base. This process was repeated for the anti-sense oligonucleotide chain with similar results, due to the palindromic nature of the DNA substrate and dimeric nature of the restriction endonucleases. In addition to these steric vdW calculations, qualitative vacuum electrostatics assessments of the protein IP and accessible surface for each structure were generated using a local protein contact potential, without solvent dielectrics, and with equilibrium charges and radii settings from the Assisted Model Building and Energy Refinement (AMBER 99) force field to evaluate IP charge complementarily to the substrate DNA (156). Given these sterics and electrostatics, the consensus participating positions from the acceptor IP were overlaid with the consensus participating residues from the donor IP to

propose a putative engineered IP on the acceptor scaffold that binds the donor substrate.

3.3.5. Engineering validations

Proposed putative engineered IP was validated both computationally and empirically. Structural mutagenesis was performed on both monomers of R.PvuII-D using a PyMol-native rotamer library (102) to generate the mutant homodimeric enzyme. Discrete mutant rotamers were auto-positioned based on calculated lowest energy, steric hindrance minimizing conformations, and then relaxed to assume along the same spatial direction as would be achieved by a natural continuous R-group. Then, in one approach, 5'-gatatc-3' DNA substrate coordinates were extracted from R.EcoRV-D, and in the other B-form substrate coordinates were generated *de novo* using nucgen (157). Each 5'-gatatc-3' substrate was then inserted into R.PvuII-D and affine space aligned along the 5'-cagctg-3' DNA substrate using the shared second A and fifth T bases as spatial and directional coordinates of reference. Upon aligning, 5'-cagctg-3' DNA substrate was deleted resulting in the R.PvuII Putative Engineered IP mutant-D, i.e. the mutant bound to the 5'-gatatc-3' DNA substrate. Hydrogen bond and polar contact patterns between the IP residues and DNA substrate involved in recognition were calculated and compared for engineered and wild-type ensembles.

Preliminary empirical validation was carried out using a cell survival assay. R.PvuII Putative Engineered IP mutant was synthesized *de novo* using a similar protocol as described in (158) and the correct synthetic sequence was confirmed by standard DNA sequencing technology. This synthetic sequence was sub-cloned into the pET-21a expression vector (Novagen), which was then transformed into *E. coli* BL21 cells. Transformed cells were cultured with appropriate selection antibiotics in the presence or absence of IPTG, which induced expression of R.PvuII Putative Engineered IP mutant.

3.4. Results

3.4.1. Objective of CF-BF

The coupled CF-BF criterion (Figure 3-3), can be used to generate a few promising candidate engineered IPs in the original scaffold that will bind to a different substrate. Upon freezing those scaffold coordinates not relevant to engineering, the CF can be calibrated to both the donor and acceptor IPs so that it can appropriately eliminate energetically impossible or unlikely residue participants, to produce some semi-engineered IP candidates that may have the desired binding ability. But given the size of many IPs, there will still be too many to fully consider energy minimization. This number can be reduced further by BF on those aspects of the candidate and other IPs

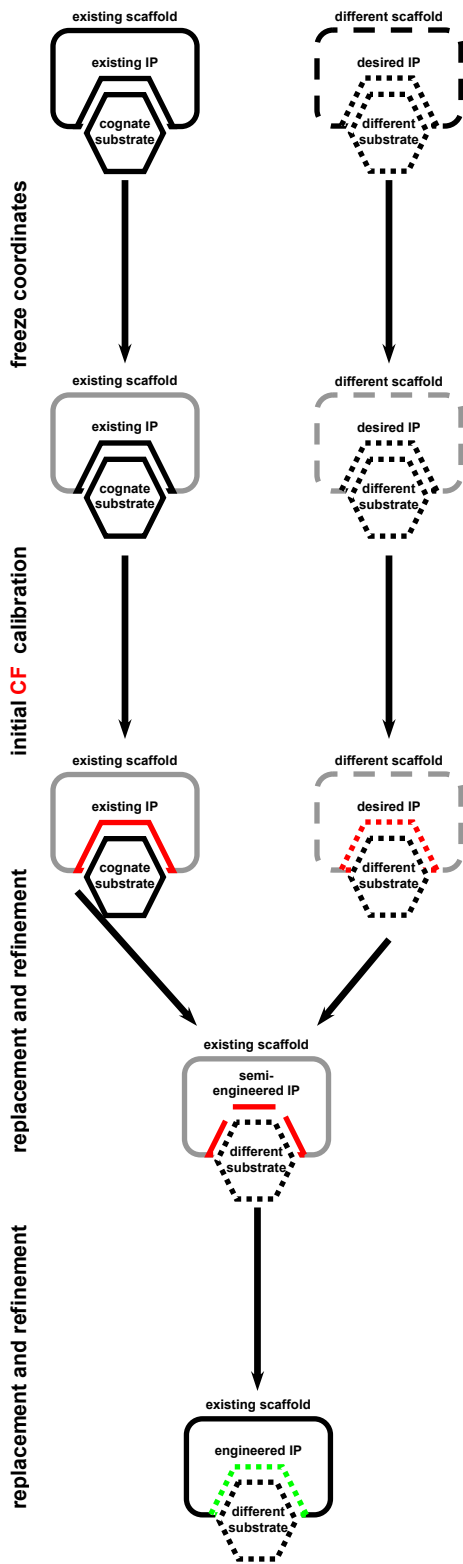


Figure 3-3: Flowchart of coupled CF-BF criterion to engineer an IP

This flowchart of the CF-BF criterion demonstrates some possibilities from Figure 3-1 to engineer an IP on the existing scaffold that binds to a non-cognate substrate than that of the existing protein. Should the CF and BF determine that the two proteins behave structurally or interfacially similar, it may be possible for one to act as an IP donor, having key residues and R-groups conformations of those residues, which can be transplanted into the other protein, the IP acceptor.

shown to be evolutionarily conserved or fit. Note that this BF permits informed selection of R^4 from a continuous set of remaining possible R-group conformations rather than discrete enumeration from a rotamer library often used in the CF performed upstream. Also note that this BF is not conditional upon CF and can be used standalone or coupled to the latter to yield results. This smaller number of engineered IPs will still be subject to downstream energy minimization as is the rest of the protein-DNA complex. However, given their reduced number, a larger portion or all of them can now be evaluated.

3.4.2. Primary Structure (PS) Biological Focusing

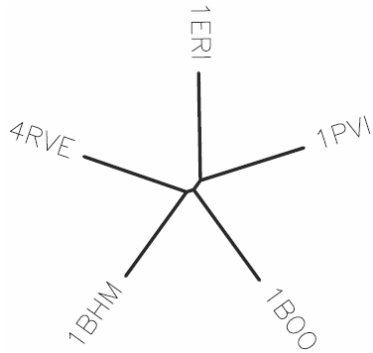
The objective of BF on PS is to focus on those candidates that have similar origins based on their polypeptide sequences. Candidate proteins are all from prokaryotic organisms (Table 3-2). Yet, it is notable that all but M.PvuII-N perform a similar biological function—sequence recognition and cleavage of DNA substrates. While a multiple sequence alignment of polypeptide sequences is not particularly revealing in Figure 3-4, the associated phylogenetic tree indicates that the methyltransferase M.PvuII-N (PDB ID: 1B00) which protects DNA from cleavage is most distant from the others. While this tree does not represent actual evolutionary patterns, it is not surprising that the monomeric M.PvuII-N may have arisen differently

Table 3-2: Origins and attributes of chosen proteins

protein	source organism	DNA substrate
R.PvuII-N	<i>Proteus vulgaris</i> (source: ATCC 13315)	5' - cag ↓ ctg -3' 3' - gtc ↑ gac -5'
R.PvuII-D	as immediately above	as immediately above
M.PvuII	as immediately above	5' - cag ^{m4} ctg -3' 3' - gtc _{m4} gac -5'
R.EcoRI-N	<i>Escherichia coli</i> RY13 (source: R.N. Yoshimori)	5' - g ↓ aattc -3' 3' - cttaa ↑ g -5'
R.EcoRI-D	as immediately above	as immediately above
R.EcoRV-N	<i>Escherichia coli</i> J62 pLG74 (source: L.I. Glatman)	5' - gat ↓ atc -3' 3' - cta ↑ tag -5'
R.EcoRV-D	as immediately above	as immediately above
R.BamHI-N	<i>Bacillus amyloliquefaciens</i> H (source: ATCC 49763)	5' - g ↓ gatcc -3' 3' - cctag ↑ g -5'
R.BamHI-D	as immediately above	as immediately above

Figure 3-4: PS properties of chosen proteins

<u>PDB ID</u>	<u>PS multiple sequence alignment</u>	
1BHM	-----MEVEKEFTTDEAKELLSKDKLIQQAYN-----	
4RVE	-----SLRSDLINALYDENQKYDVCGIISAEGK-----	
1BOO	MLNFGKKPAYTTSNGSMYIGDSLELLESFPEESISLVMTSPPFALQRKKEYGNLEQHEYV	
1ERI	---SNKKQSNRLTEQHKLSSQGVIGIFGDYAKAHD LAVGEVSKLVKKALSNEY PQLSFRYR	
1PVI	-----MSHPDLNKLLELWPHIQEYQDLALKHG-----	
	:	
1BHM	-----EVKTSICSPIWPATSKTFTTINNTEKNCNGVVPikelCYT---LLEDTYNW	
4RVE	-----IYPLGSDTKVLSTIFELFSRPIINKIAEKHGVIIEEPKQONHYPDFTLYKPSPEP	
1BOO	DWFLSFAKVVNKKLKP DGSFVDFGGAYMKGVPARSIYNFRVLIRMIDEVGFFLAEDFYW	
1ERI	DS----IKKTEINEALKKIDPDLGGTLFVSNSSIKPDGGIVEVKDDYGEWRVVLVAEAKH	
1PVI	-----INDIFQDNNGKLLQVLLITGLTVLPGREGNDAVDN-----AGQE	
	.	
1BHM	YREKPLDIL--KLEKKKGG----PIDVYKEFIE-----NSELKRVG---	
4RVE	NKKIAIDIK--TTYTNKENEKIKFTLGGYTSFIR-----NNTKNIVYPPFD	
1BOO	FNPSKLPSP--IEWVNKRKIRVKDAVNTVWWFSKTEWPKSDITKVLAPYSDRMKKLIEDP	
1ERI	QGKDIINIRNGLLVGKRGDQDLMAAGNAIERSHKN-----ISEIANFMLSSES	
1PVI	YELKSIDID-----LTKGFSTHHH-----MNPVIAIKY	
	:	
1BHM	MEFETGNISSAHRSMNKL LLLGLKHGEI-DLAIILMPIKQLAYYLTDRVTNFELEP----	
4RVE	QYIAHWIIGYVYTRVATRKS SLKTYNINELNEIPKPYKGVKVF LQDKWVIAGDLAGSGNT	
1BOO	DKFYTPKTRPSGHDIGKSF SKDNNGSIPP NLLQISNS ESNGQYLANCKLMGIKAHPARFP	
1ERI	HFPYVLFLEGSNFLTENISITRPDGRVVNLEYN SGILNRLDRLTAANYGMPINSNLCINK	
1PVI	RQVPWIFAIYRGIAIEAIYRLEPK----DLEFYDKWERK-----	
	:	
1BHM	-----YFE-----LTEGQPFIFIGFNAAEAYNSNVPLIPKGS DG	
4RVE	TNIGSIHAHYKD-----FVEGKGFIDSEDEF LDYWRNYERTS QLRND	
1BOO	AKLPEFFIRMLTEPDDLVDIFGGSNTTGLVAERESRKWISFEMKPEYVAASAFRFLDNN	
1ERI	FVNHKDKSIMLQAAS-----IYTQGDGREWDSKIMFEIMFDISTTSLRVLG	
1PVI	-----WYSDGHKDINNPKIPVKYVMEHGTKIY----	
	:	
1BHM	MSKRSIKKWKDKVENK-----	Consensus key
4RVE	KYN-NISEYRNWIYRGRK-----	* - single, fully conserved residue
1BOO	ISEEKITDIYNRILNGESLDLNSII	: - conservation of strong groups
1ERI	-----RDLFEQLTSK-----	. - conservation of weak groups
1PVI	-----	- no consensus



The primary structures of the chosen proteins were compared using multiple sequence alignment (top panel) and associated phylogenetic tree (bottom panel). Adapted from (159).

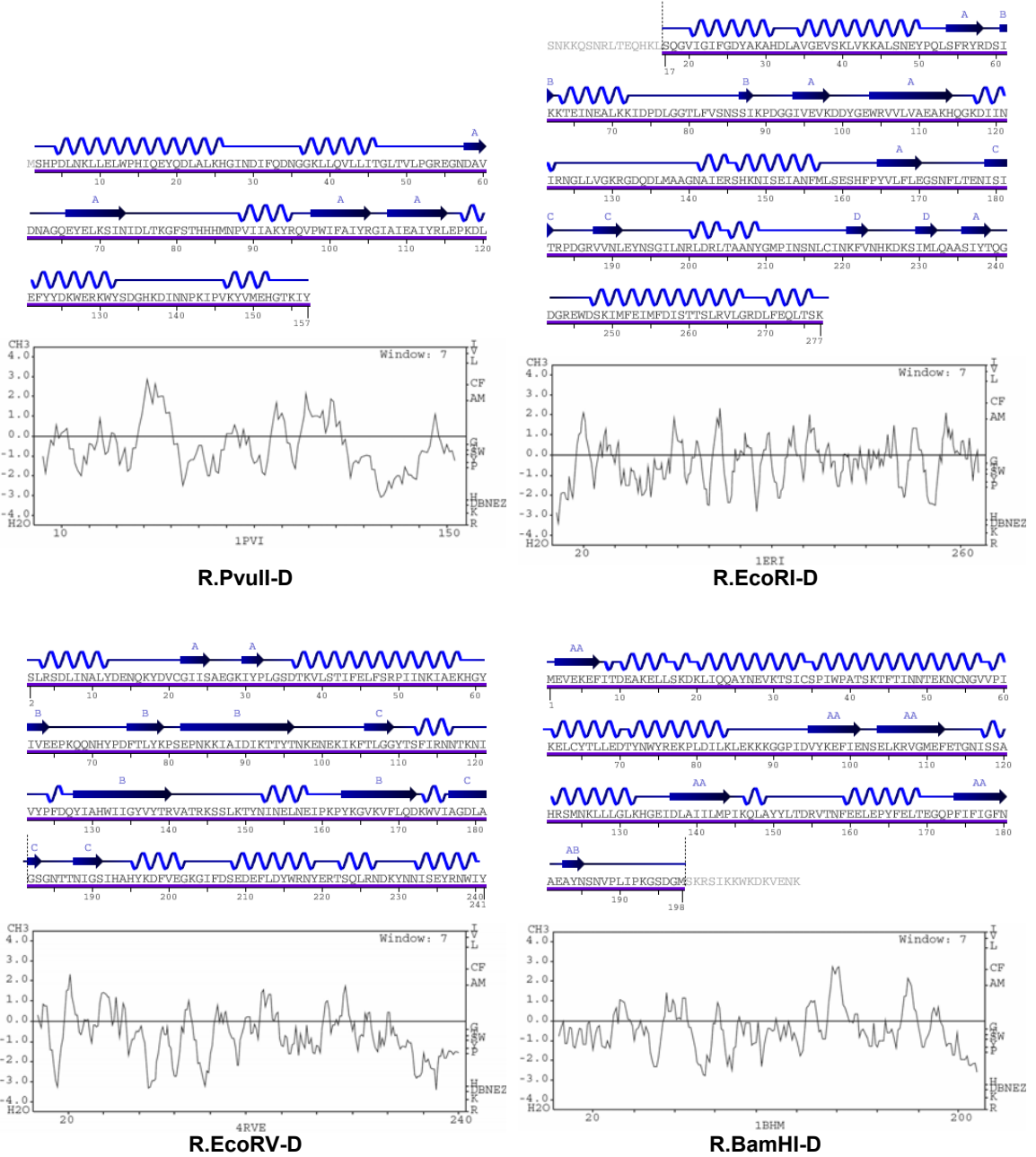
than the dimeric restriction endonucleases. Thus, this PS BF deemphasizes the IP of this methyltransferase as a source of donor residues and R-group conformations for R⁴ into the R.PvuII-N/D dimer.

3.4.3. Secondary Structure (SS) and Hydropathy BF

The objective of BF on SS is to focus on those remaining candidates that have similar local hydropathic profiles and conformations of the IP polypeptide BB, such as the alpha helix and beta sheet. In addition, given evolutionary fold conservation at protein IPs, such as active and allosteric sites, it may be worthwhile to compare how these local conformations interact with the DNA substrate. Also, this conservation may influence the choices made in R⁴ since certain residues are more capable in participating in particular SS, as they are able to adopt the necessary backbone dihedral angles. A mapping of these dihedral angles to the corresponding SS and capable residues can be found in a Ramachandran plot.

For this analysis, hydropathic profiles remained uninformative, but similarity in secondary structure motif interactions to DNA grooves permitted further focusing (Supplementary Figure 3-2). It was calculated that both R.EcoRI-N/D and R.BamHI-N/D tend to have a greater proportion and longer stretches of alpha helices (shown as waves) than beta sheets (shown as arrows), while both R.EcoRV-N/D and R.PvuII-N/D are more

Supplementary Figure 3-2: Secondary structure and hydrophathy properties of remaining proteins after primary structure focusing



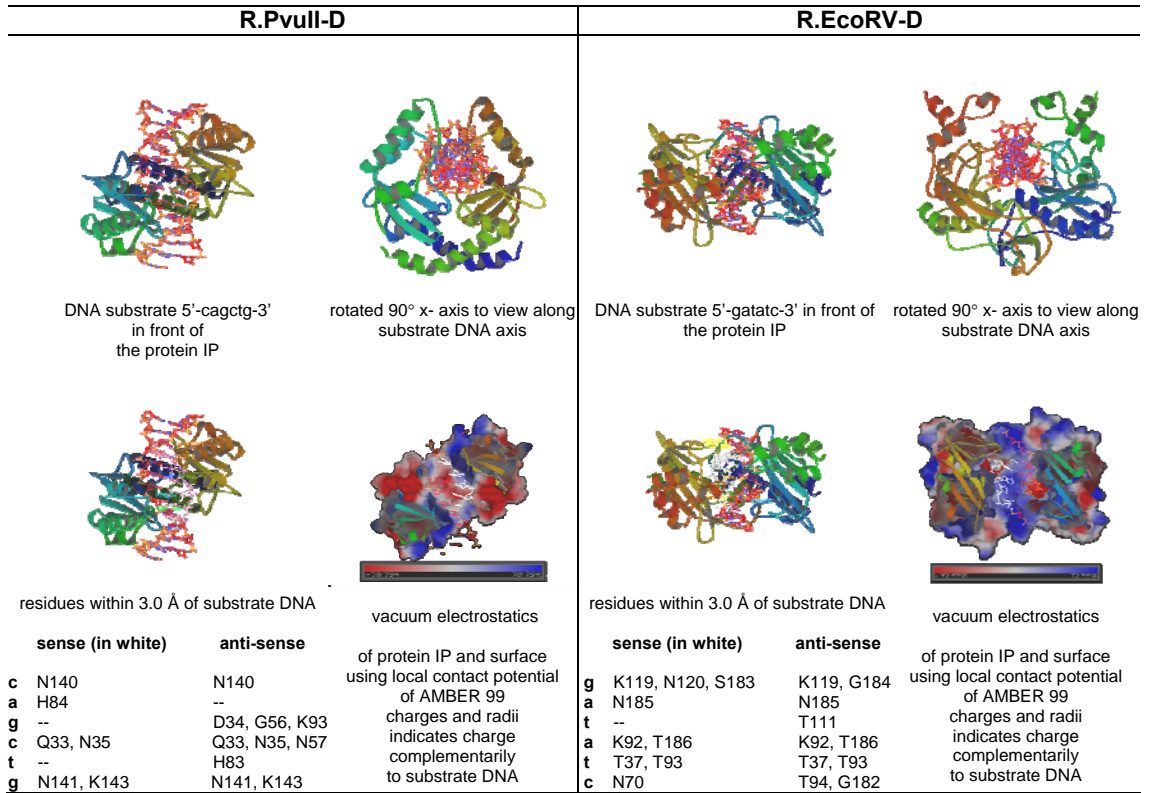
balanced in beta sheet content. Notably, both R.EcoRI-N/D and R.BamHI-N/D approach and recognize the DNA substrate from the major groove via an alpha helix and a loop and produce 5' sticky ends, while both R.EcoRV-N/D and R.PvuII-N/D do so from the minor groove via a beta sheet and beta-like turn and produce blunt or 3' sticky ends. Thus, this SS BF deemphasizes R.EcoRI-N/D and R.BamHI-N/D as sources for IP donation to R.PvuII-N/D, leaving R.EcoRV-N/D as a more biologically promising candidate.

3.4.4. Tertiary (TS) and Quaternary (QS) Structure BF

The objective of BF on TS and QS is to spatially align the remaining two well-focused candidates and readily identify the IP residues close enough to interact with the substrate. Given that both R.PvuII-N and R.EcoRV-N recognize and bind to a uniform, helical substrate, this can facilitate R⁴ by acting as a common coordinate reference from which residues from the donor IP can be mapped onto positions on the acceptor IP. A promising mapping can be confirmed using qualitative vacuum electrostatics assessments of the protein IP and accessible surface, since the engineered IP should have the electrostatic profile of the donor IP while the rest of the scaffold should remain as-is.

Longitudinal and axial views of the DNA substrate were presented against each protein for orientation purposes (upper panels of Figure 3-5).

Figure 3-5: Tertiary and quaternary structure properties of remaining proteins after primary and secondary structure focusing



From the longitudinal views, some asymmetric distortion of the DNA substrate upon binding can be seen. This may contribute to the lack of identical IP residues found within the 3 Å boundaries for both the sense and anti-sense strands of the DNA substrate. Upon selecting a DNA strand as points of origin, which IP residues within this 3 Å boundary are closest to which DNA base are explicitly indicated (lower panels of Figure 3-5). Taking advantage of the existing symmetries, the consensus IP position 140 on R.PvuII-D was determined to interact with the first base of the DNA substrate, and so on. Similar symmetries showed a consensus IP residue LYS on R.EcoRV-D to interact with the first base of the DNA substrate. By integrating discrete levels of biological abstraction, from PS to SS to TS and QS, in a systematic manner, a putative engineered IP is mapped on the original R.PvuII-N scaffold (Figure 3-6) to confer specificity and bind the R.EcoRV 5'-gatatc-3' substrate.

3.4.5. Computational validation via structural mutagenesis

The putative engineered IP mapping was validated structurally using relaxed rotamer-based mutagenesis and hydrogen bond and polar contact pattern comparison. Discrete mutant rotamers were positioned with greater than majority occupancy at that lowest energy with little to no steric hindrance. After positioning, local relaxing permitted the mutant to assume

Figure 3-6: BF for putative engineered IP on original R.PvuII scaffold to bind R.EcoRV 5'-gatatc-3' substrate

R.PvuII Putative Engineered IP mutant		
non-cognate DNA bases	residues within 3.0 Å of substrate DNA	
	sense	anti-sense
g	140K	140K
a	84N	84N
t	93T	93T
a	33K, 35T	33K, 35T
t	83T	83T
c	141N, 143T	141N, 143T

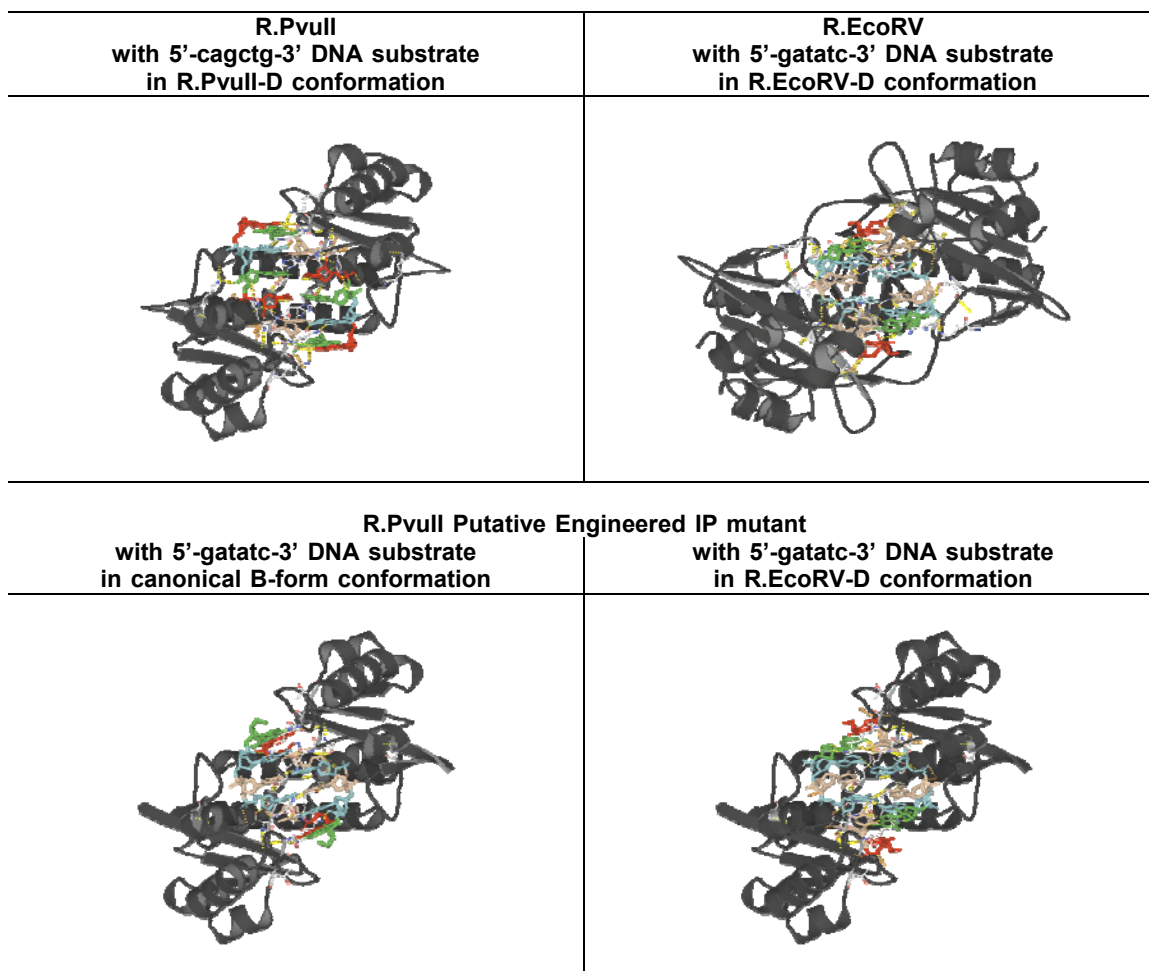
having consensus residues of R.EcoRV to bind 5'-gatatc-3' substrate,
at consensus 3.0 Å positions of R.PvuII

similar spatial direction as wild-type residue had, thus achieving continuous R-group positioning. Affine space alignment repositioned the 5'-gatatc-3' DNA substrate from the R.EcoRV-D orthogonal coordinate system to that of the R.PvuII Putative Engineered IP mutant-D ensemble and then aligned it to the 5'-cagctg-3' DNA substrate. Though the R.EcoRV-D 5'-gatatc-3' DNA substrate was slightly more bent, and the B-form one slightly less, than the R.PvuII-D 5'-cagctg-3' DNA substrate, the affine space alignments achieved good fits with expected deviations occurring at the substrate extremes, with root mean square deviations (RMSD) of 2.02 Å and 2.07 Å for the R.EcoRV-D and canonical B-form conformations, respectively. Hydrogen bonding and polar contacts were made by the mutant' IP residues to all bases and BB in both conformations of 5'-gatatc-3' substrate, just as for the wild-type ensembles (Figure 3-7). Thus, this suggests that the mutant is labile yet specific enough to approach and make recognition contacts with B-form DNA, and then maintain these contacts while bending it like R.EcoRV-D to expose the scissile phosphate for enzymatic cleavage. Computationally, this R.PvuII Putative Engineered IP mutant exhibited promising recognition characteristics and was further evaluated empirically.

3.4.6. Preliminary experimental validation

A cell survival assay was performed to determine whether the

Figure 3-7: Computational validation based on hydrogen bond and polar contacts with respect to steric hindrance patterns



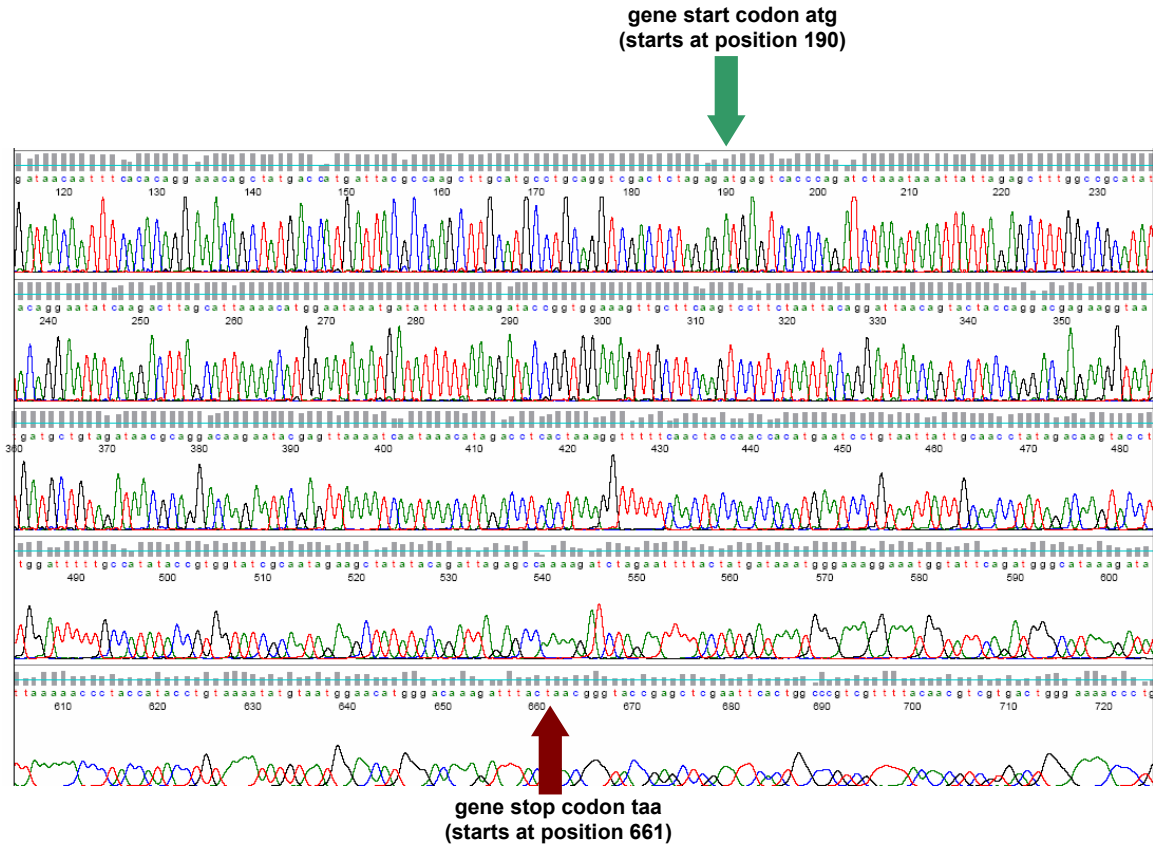
Wild-type R.PvuII-D interacts with DNA substrate 5'-cagctg-3' (top left), R.EcoRV-D with 5'-gatatc-3' (top right), and R.PvuII Putative Engineered IP mutant also with canonical B-DNA-like (bottom left) and R.EcoRV-D-like (bottom right) 5'-gatatc-3' in order to make hydrogen bond and polar contacts while avoiding steric hindrance. DNA are color coded according to bases (alanine = aquamarine, cytosine = crimson red, guanine = green, thymine = tan), IP residues according to standard Corey, Pauling, Koltun atom colors (carbon = white, oxygen = red, nitrogen = blue), and hydrogen bond and polar contacts highlighted in dotted lines (contacts from IP residues to any other atoms in yellow, from DNA 6-mer to any other atoms in orange).

engineered R.PvuII has any enzymatic activity in cutting DNA. A synthetic R.PvuII Putative Engineered IP mutant gene was cloned into pET-21a expression vector, validated by DNA sequencing (Supplementary Figure 3-3), and cultured on agar plate in the presence or absence of isopropyl- β -D-thiogalactoside (IPTG). Without IPTG, the cells grew normally and formed colonies; with IPTG, cells did not grow and no colonies were found on the plate (Supplementary Figure 3-4). The result of this cell survival assay suggested that the expression of the R.PvuII Putative Engineered IP mutant in *E. coli* lead to cell death presumably due to digestion of the host chromosomal DNA (160).

3.5. Discussion

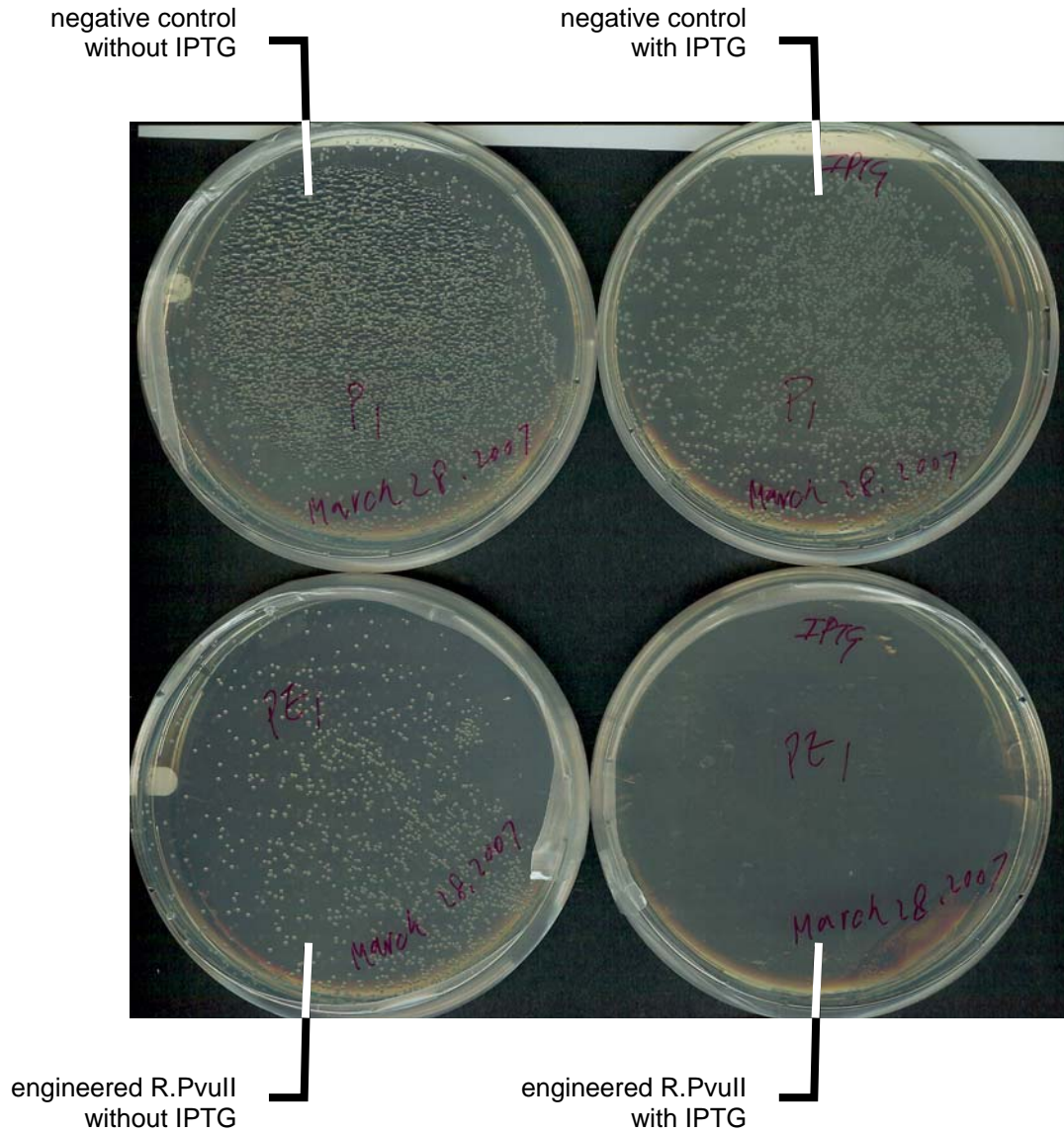
While CF-BF produced a promising putative engineered IP, the computational and empirical validations confirmed its properties. The computational validation of R.PvuII-N scaffold with the engineered IP examined the hydrogen bonding and polar contact patterns to show they, like the R.EcoRV-D IP, interact with all the bases in the substrate 5'-gatatc-3' DNA. Preliminary empirical validation using the cell survival assay suggested that the R.PvuII Putative Engineered IP has enzymatic activity in cutting DNA. The engineered specificity was examined further biochemical assays, such as an enzymatic function assay (Supplementary Figure 3-5).

Supplementary Figure 3-3: Synthetic R.PvuII gene validated by sequencing

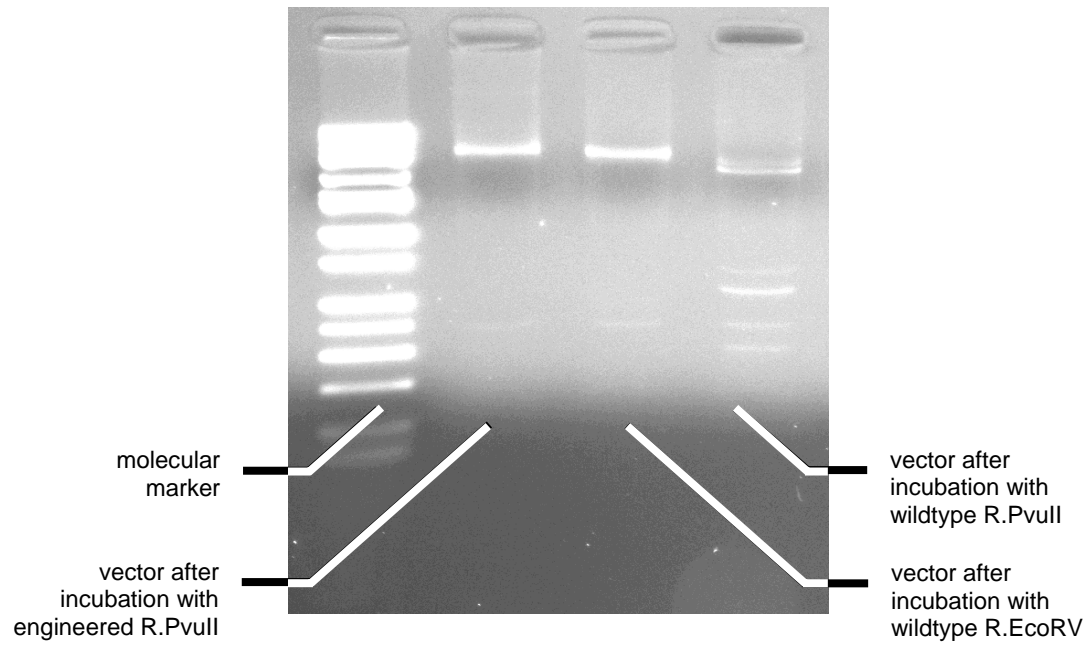


The synthetic R.PvuII gene sequence was verified to be accurate, with no insertions or deletions (indels) by sequencing using a 3730xl DNA Analyzer (Applied Biosystems, CA, USA).

Supplementary Figure 3-4: *In vivo* cell survival assay for R.PvuII Putative Engineered IP mutant in *E. coli*



Supplementary Figure 3-5: *In vitro* enzymatic function assay for R.PvuII Putative Engineered IP mutant in *E. coli*



These validations, in turn, will inform and improve the CF-BF criterion through further iterations of this process for these types of nucleic acid binding enzymes.

The IP engineering possibilities are nearly as vast as the diversity of biology itself. Using CF alone, R^4 of a 6-mer DNA recognition site and a monomer of a dimer restriction enzyme, assuming that there is one residue interacting with each base of DNA and a limited number of rotamers in a library represents all possible R-group conformations of the twenty naturally-occurring residues, would have required the evaluation of possible IP sequences equal to the size of the rotamer library to the sixth power in number. The CF-BF reduces this to a subset of the twenty naturally occurring residues (the subset being the ten polar and charged residues in this restriction endonuclease experiment herein) and R-group conformations, which are known to participate in the IP and interact with a similarly structured substrate.

The benefits of successful IP engineering are equally numerous. Engineered IPs may lead to programmable proteins, such as restriction endonucleases that not only act as research tools, by enabling targeting, mapping and manipulation of genes and genomes, but as clinical technologies as well, by facilitating cleaving out a disease gene and repairing it with a working version *in vivo*.

4. Multiple-conformation engineering of nucleic acid binding proteins

*The overall struggle for existence
of living beings is therefore
not a struggle for raw materials—
the raw materials of all organisms are available
in excess in the air, water, and ground—
nor for energy, which in the form of heat
is plentiful in every body,
but rather a struggle for entropy,
which becomes available in the flow of energy
from the hot sun to the cold earth.*

– Ludwig E. Boltzmann
Austrian physicist

This chapter has been adapted from a manuscript that was joint work with Qihai Wang, Ivelin Georgiev, Bruce R. Donald, Jingdong Tian:

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Automated and accurate engineering of a superior nucleic acid enzyme. *In revision.*

and partially from a research meeting abstract that was joint work with Qihai Wang, Ivelin Georgiev, Bruce R. Donald, Jingdong Tian:

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Molecular ensemble engineering and evaluation for targeted genome therapeutics. *Biomedical Engineering Society 2009 Annual Meeting*. 2009, Pittsburgh, PA.

4.1. Motivation

We report the first molecular engineering protocol which integrates the deterministic algorithms minimized side-chain Dead-End Elimination (minDEE), ordered conformational enumeration (A*), and ensemble-based scoring (K*) for scanning and redesigning protein-nucleic acids interactions. Furthermore, we report application of this protocol to the scanning and redesign of a Type II restriction endonuclease, R.PvuII, with predictions validated by *de novo* gene synthesis, cell-free protein expression and enzyme function assays.

This protocol accurately scanned and characterized R.PvuII residues for mutation tolerance. Redesign of a tolerant residue using this protocol predicted 8 mutants to bind cognate DNA substrate, of which the top 6 showed enzymatic activities. The computed ensemble-based binding affinity scores, rather than single-conformation, or more specifically global minimum energy conformation (GMEC)-based bound energy values, correlated better with experimental results. Remarkably, the top-scoring redesign exhibited significantly faster activity, reducing the DNA digestion half-time to a third as compared to the wildtype while maintaining substrate specificity. These results demonstrated the utility of this protocol in automated and accurate ensemble-based engineering of protein-nucleic acids interactions.

4.2. Overview

Interactions between proteins and nucleic acids are ubiquitous and essential for the transduction of biochemical information into physiological action in nature, basic research, and translational medicine (5,6,161-169). Accurate engineering of these interactions are of keen importance and have been attempted through both experimental and computational means. Experimental strategies generally involved random diversification, selection and amplification (77,138,170-172). Computational redesign to alter DNA recognition specificity of a homing endonuclease has been demonstrated on single conformations, rather than ensembles, using stochastic searches, with specificity further improved by directed evolution experiments (65). Other rational DNA enzyme redesigns have used heuristic approaches based on sequence homologies (173), or made compatible modular fusions between existing proteins (60,61,174) or between proteins and DNA (175,176).

Although significant progress has been made, molecular engineering of protein-nucleic acids interactions through computational design remains exceedingly challenging (53,177). Natural proteins have evolved over long periods of time by balancing properties such as stability, solubility, substrate affinity, specificity, and catalytic activity. Accurately and consistently engineering mutants with equal or better properties while maintaining such balances is a significant challenge. This challenge can be highlighted using examples of restriction endonucleases (REases), which are indispensable in

nature and the laboratory for binding and cutting specific DNA sequences (16). Most REases have tightly-coupled binding and catalytic activities (178). Given these coupled interactions, engineering improved or new properties without compromising others that are already satisfactory is difficult.

Consider the Type II REase R.PvuII as an example. R.PvuII is isolated from the Gram-negative human pathogen *Proteus vulgaris*. It is the smallest known REase with only 157 amino acid residues (18.3 kDa) per monomer, and its enzymatic activity can be easily abolished by single point mutations (41,179). The functional enzyme is a homodimer that binds double-stranded palindromic DNA substrate 5'-cagctg-3'. Each monomer is comprised of three structural regions: a subunit interface region, a catalytic region, and a DNA recognition region (41). The N-terminal 46 residues form the subunit interface region through two alpha helices connected by a loop. The subsequent residues form interspersed catalytic and the DNA recognition regions in primary sequence. The catalytic region consists of mixed beta sheets and an alpha helix, which allows coordinated access and action upon the DNA substrate. The recognition region comprises two subregions separated in primary sequence, each containing one helix and one short loop. The two short loops (residues 80-84 and 140-144) make direct and specific interactions with DNA substrate bases. With Mg^{2+} also coordinated, this cooperative association proceeds to chemically-coordinated catalytic

cleavage of both DNA scissile phosphates between +1 and -1 base pairs (BPs) from the substrate midpoint, yielding blunt-ended DNA restriction sequences 5'-cag-3' and 5'-ctg-3'. Replacing Mg^{2+} with Ca^{2+} permits binding and unbinding but not subsequent cleavage of the DNA substrate (44,180). These sequence-structure-function characteristics make R.PvuII an attractive, albeit extremely challenging engineering example.

To accurately and efficiently engineer protein-nucleic acid interactions, we have made significant advances over prior works (112,158,159,181) and report here the development of a deterministic and automated computational protocol and an efficient experimental validation pipeline. In addition, we report the successful application towards accurately engineering mutants for activity comparable to or faster than the R.PvuII found in nature.

4.3. Methods

4.3.1. Scanning NABP residues for mutation tolerance

The protocol was applied to scan residues for mutation tolerance, as a function of computed binding energy loss. For scanning, each position on the mutation map on both or one of the two monomers was allowed to mutate to the wildtype or alanine residue by setting the sequence space of allowable residues accordingly. At each position all rotamers of wildtype residue or alanine, which has one conformation, were placed. The homodimer target volumes for the residues being scanned were S81 137.78 Å³ and N140 or

N141 176.34 Å³ and volume windows were set to 100% of these volumes. Entire DNA-REase complex and entire REase were used as bound and unbound models, respectively, to permit cross-comparison of binding energies and affinities among scans of different residue positions.

4.3.2. Redesigning NABP residues for functional mutations

The protocol was applied to redesign mutation-tolerant residues for binding the DNA substrate, 5'-cagctg-3'. For redesign, each position on the mutation map was allowed to mutate to the wildtype or 18 other possible non-cyclic, naturally occurring residues by setting the sequence space of allowable residues accordingly. The homodimer target volumes for the residues being scanned were chosen similarly, but the volume windows were set to 20% of these volumes. The number of residues in the bound model was reduced to a steric shell of all atoms within 8.0 Å from any atom belonging to the -3c::+3g BP and then extended to include whole residues of such atoms. All residues having one atom within 3.0 Å from any atom belonging to the mutable residue was permitted to be flexible. The same REase residues in the unbound model were reduced to those in the steric shelled bound model. These steric shelled models of the entire DNA-REase and entire REase facilitated rotameric mutations to many allowable residue types while sterically constraining such mutations to allowable conformations.

4.3.3. Engineering validations

Wildtype R.PvuII gene and its mutants were *de novo* synthesized. The PURExpress *In Vitro* Protein Synthesis Kit (New England Biolabs, MA, USA) was used for expressing wildtype and mutant R.PvuII proteins following manufacturer instructions. Each PURExpress reaction was incubated at 37 °C for 3 hours. The amount of proteins synthesized was monitored by Western blot analysis using anti-His6-tag antibody and Amersham ECL Plus Western Blotting Detection System (GE Healthcare). The enzyme activity assays were carried out in 20 µl reactions containing a mixture of 1 µl protein product, 150 ng of DNA substrate and 2 µl of NEBuffer 3. After incubating for specified amounts of time (0-120 min) in a 37°C water bath, reactions were immediately stopped by adding EDTA (10 mM) and gel loading dye and stored on ice until analysis. When all reactions were completed, 20 µl of each reaction was analyzed by agarose gel electrophoresis (1% agarose in TAE buffer 0.5 ug/ml ethidium bromide). Gel images were taken using an Alpha Innotech gel documentation system and the densitometry of DNA bands were determined using the averaged integrated density value method of AlphaEaseFC software (Alpha Innotech, version 4.0.0). Replicates of three (n=3) were performed of the above procedure. Half-time was computed for mutants that completely digested the DNA substrate (3113 bp) to +cut DNA (2514 bp) within 2 hours.

4.4. Results

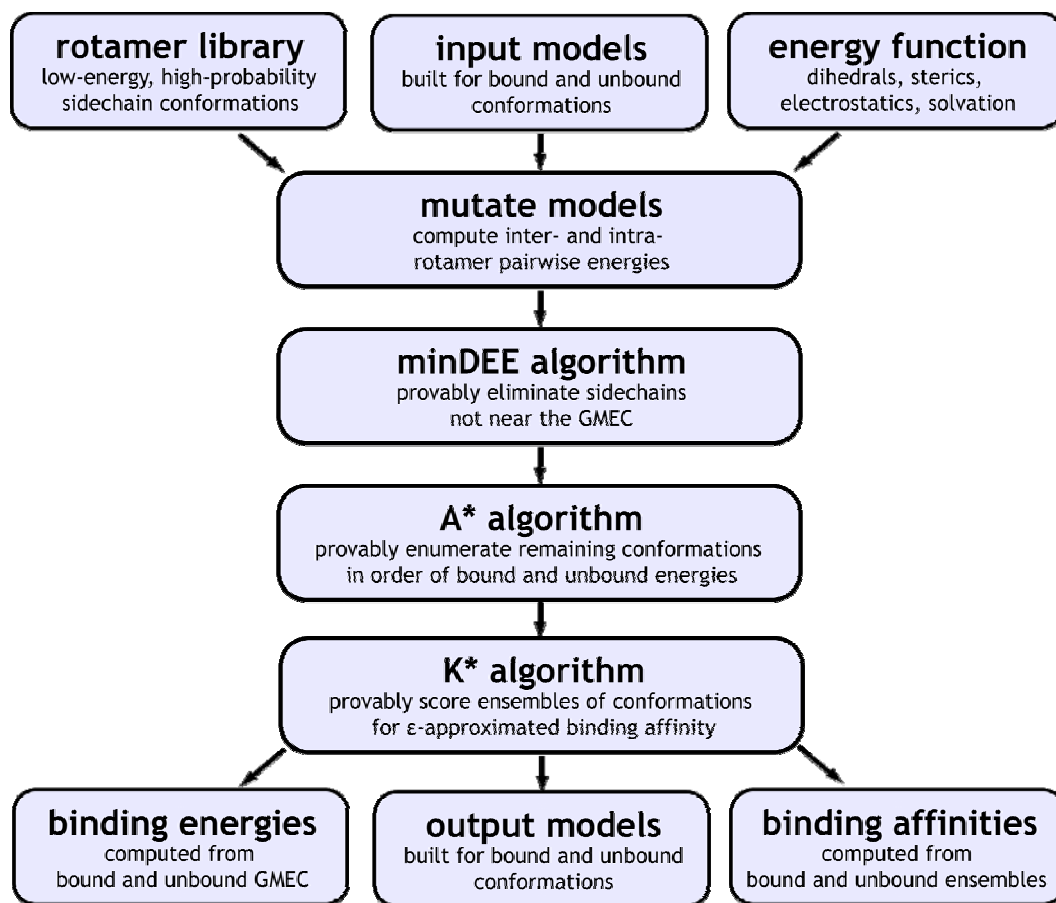
4.4.1. Molecular engineering protocol for NABPs

The protocol we developed started with building 3D models from structures as input, followed by mutating and pruning the mutations with the deterministic algorithms 1) minimized side-chain Dead-End Elimination (minDEE), 2) ordered conformational enumeration (A*), and 3) ensemble-based scoring (K*), and concluded by computing bound and unbound conformational energies, binding energies, binding affinities, and building 3D structure models as output (Figure 4-1). We also implemented a corresponding experimental validation protocol for the mutant proteins (Supplementary Figure 4-1). The input 3D models were built from atomic coordinates of crystallographic structures bound and unbound to nucleic acids. The algorithms then mutated these models using a rotamer library of low-energy, high-probability side-chain conformations and computed an array of minimum and maximum bounds and ranges on pairwise energies using an energy function composed of dihedral, steric, electrostatic, and solvation pairwise-decomposable terms. The conformations were then provably pruned, enumerated, and scored by the minDEE (112), A* (115), and K* (112,124) algorithms, respectively. The algorithms computed 3D structures of all unpruned bound and unbound conformations, as well as their energies. Three means of scoring were used in this study: conformational energy, binding energy, and binding affinity.

Input models. The bound and unbound 3D structure input models were built from source crystallographic structure coordinates for the *holo* form phased at 2.60 Å (41) (Figure 4-2A). In this process, heavy atoms for REase or DNA were retained and others discarded; broken or missing residues were repaired by remodeling to the extent permitted by the electron density; hydrogen atoms were computationally added and the hydrogen bonding network mapped; and N, Q, and H residue side-chains were flipped in order to optimize geometric and biochemical properties. This bonding network mapping revealed that the -3c::+3g (and through symmetries, the +3g::-3c) BP in the substrate center interacted exclusively with residues S81, N140, and N141 of one REase monomer or the other (Figure 4-2B). Although these REase residues were located on flexible loop regions, the bonding network and packing stabilized the interactions among each other and with the DNA substrate. These hydrogen bonding and packing interactions were automatically considered by the algorithms in the protocol when they were subjected to this example.

Input models building. Input models of R.*PvuII* bound and unbound to cognate DNA were built from X-ray crystallography determined atomic structure data resolved at 2.60 Å (PDB accession ID: 1PVI (41)) from the

Figure 4-1: Molecular engineering protocol for computing ensemble-based binding energies and affinities of protein-nucleic acids interactions



Supplementary Figure 4-1: Experimental protocol for validating mutant proteins

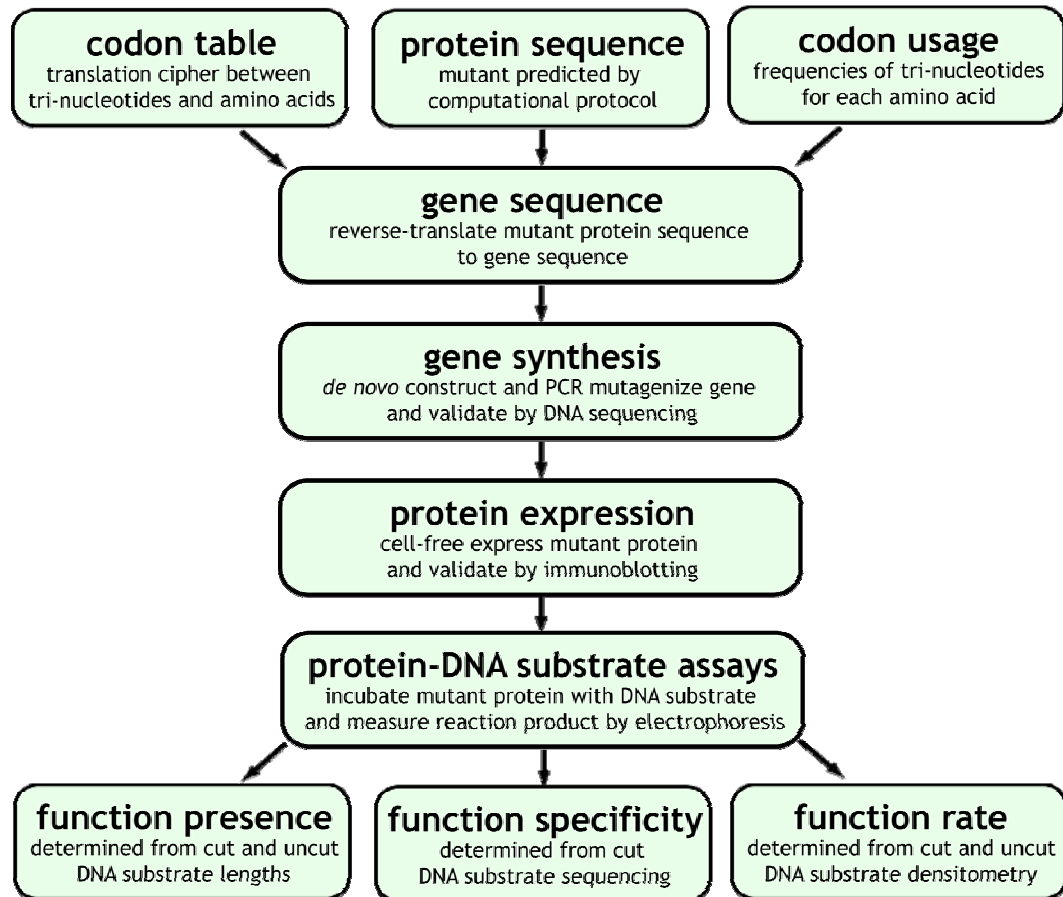
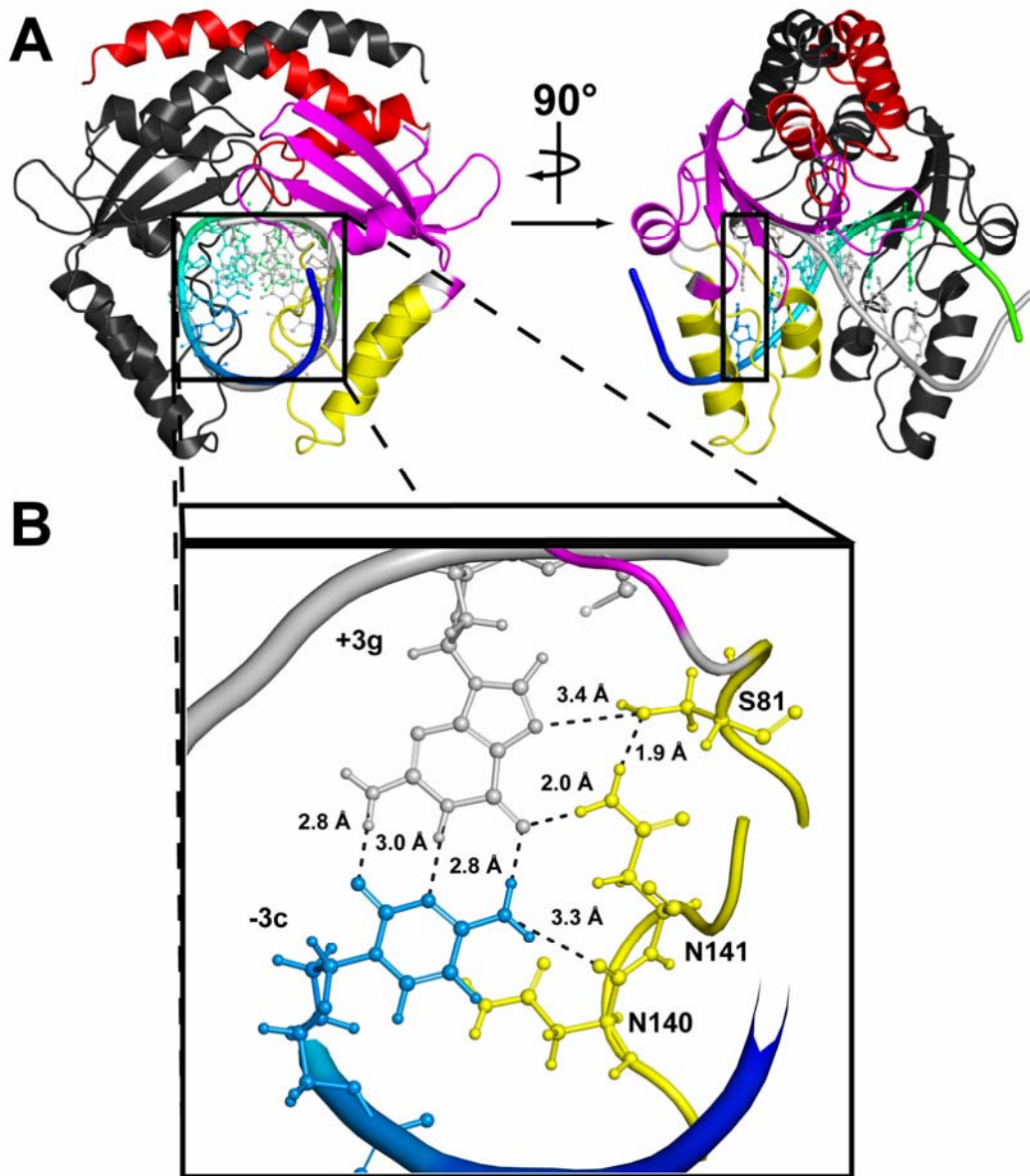


Figure 4-2: Input structure model of R.PvuII with interactions and residues of interest



(A) Crystallographic structure of R.PvuII bound to cognate DNA sequence 5'-cagctg-3'. One monomer of protein dimer is colored by subunit interface (red), catalytic (magenta), and recognition (yellow) regions, while the other is monochrome; one strand of DNA double strand is colored from 5'- (blue) to 3'- (cyan) termini, while the other is monochrome. (B) Magnified area of built model indicates residues and interactions of interest for computational engineering.

Protein Data Bank (32). Each monomer subunit of the crystallized homodimer was built and refined independently (41). Heavy atom entries belonging to the RE or DNA were retained while the rest, such as those for explicit water molecules, were discarded. All residue numbers were reassigned from matching numbers on unique chains to unique numbers throughout all RE and DNA chains in the entire bound complex. Wildtype residues Y94 on both monomers were remodeled as A94 due to insufficient resolution in the electron density beyond the C_β atoms in the side-chain. Protonation and hydrogen bonding were incorporated into the built model. At 2.60 Å resolution, heavy atoms lacked protonation and were supplemented computationally using Reduce (95). Side-chain ring moieties for N, Q, and H residues were inspected after protonation and flipped 180°, when necessary, to optimize the hydrogen bonding network with MolProbity (97,98). Nomenclature used for model constituents were made to conform to Assisted Model Building and Energy Refinement (AMBER) template conventions (39). Given model evaluation at physiological pH, all histidine residue imidazole side-chains were inspected and reassigned to HID, HIE, or HIP based on protonation of N_δ, N_ε, or both N_δ and N_ε atoms, respectively. The unbound model was obtained by removing DNA residues from the bound model. Both bound and unbound models were used in their entirety for scanning and sterically shelled to an 8.0 Å region around the -3c::+3g BP for redesign.

Model mutation. The algorithm mutated models according to a sequence space of allowable residues at each position of mutation maps designated from examination of the literature (41,43,159) and mapping of the native hydrogen bonding network. Maps for interactions with the -3c::+3g BP included S81, N140 and N141 residues. The allowable residues were set to either wildtype and alanine or all 19 types according to whether scanning or redesign was performed, respectively. Side-chain conformation of the mutant residue was initially placed similarly to those of the wildtype residue and rotated according to idealized and experimentally observed dihedral angle geometries specified in the rotamer library (103). Initial rotamer placements were later energy minimized with bounds using a steepest-descent-based approach.

Pairwise energy computation. Inter- and intra- rotamer energies, and corresponding rotamer voxel energy maxima, minima, and ranges, were computed using a composite energy function and stored in a pairwise energy matrix for expedient retrieval in subsequent protocol stages. Template energy of the C_α backbone was computed once, since it would remain unchanged and contribute a constant energy for all side-chain mutations. Pairwise components in the computation were categorized as DNA substrate, RE steric shell, and RE active site residues. The RE steric shell consisted of RE residues not part of the mutation map while the RE active site residues

were those that were part of the mutation map. Pairwise energies were computed between DNA substrate and RE steric shell, DNA substrate and RE active site, RE steric shell and RE active site, and between RE active site residues. Singleton energies for intra-residue contributions were also computed for DNA substrate, RE steric shell, and RE active site residues. These energies were subject to change as rotamers were permitted to energy minimize. The energy function was composed of linear summation of pairwise decomposable terms for AMBER bonded dihedral angles interaction, non-bonded vdW steric interaction, and non-bonded electrostatic Coulombic interaction terms for explicit DNA and RE atoms as well as a term for Lazaridis-Karplus EEF1 implicit solvation (111), of the form:

$$\begin{aligned}
E &= E_{\text{bonded dihedral angles}} \\
&+ E_{\text{non-bonded vdW sterics}} \\
&+ E_{\text{non-bonded Coulombic electrostatic}} \\
&+ E_{\text{Lazaridis-Karplus implicit solvation}} \\
E &= \sum_{\text{dihedral angles}} \frac{V_n}{2} [1 + \cos(\eta\phi - \gamma)] \\
&+ \sum_{i < j}^{\text{atoms}} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{12}} \right] \\
&+ \sum_{i < j}^{\text{atoms}} \frac{q_i q_j}{\epsilon R_{ij}} \\
&+ \Delta G_i^{\text{reference}} - \sum_{j \neq i} f_i(R_{ij}) V_j
\end{aligned}$$

with variables defined similarly as (39). The AMBER charge model and vdW parameters, rather than an explicit 10-12 term, represented hydrogen bonds.

The implicit solvation model was suitable, given crystallographic structures for all available R.PvuII bound to DNA (PDB accession IDs: 1PVI resolved at 2.60 Å (41), 2PVI resolved at 1.76 Å (42), 3PVI resolved at 1.59 Å (43), 1EYU resolved at 1.78 Å (44), 1F0O resolved at 2.50 Å (44)) were found to be lacking explicit water molecules between the -3c::+3g BP and aforementioned mutation map residues. Parameters for the energy function were set according to successful evaluations between proteins and amino acid substrates (114), as follows. A steric threshold was applied so that atomic vdW radii overlap by 1.5 Å or less are allowed prior to minimization, with overlaps greater than this deemed too great a clash to relax away from using minimization. A dielectric constant of 6.0 was applied with dielectric effects scaled by distance. The solvation energies effects were similarly scaled by a multiplicative factor of 0.05. Electrostatic energies were computed for heavy and hydrogen atoms.

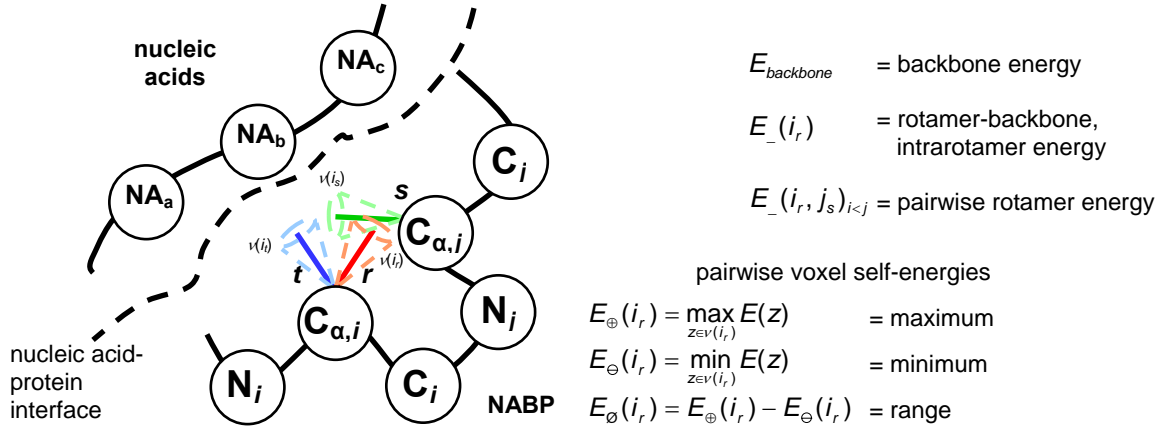
Model ensemble generation and pruning. After building the bound and unbound input structures, the algorithm generated bound and unbound conformations by building side-chain conformations using the rotamer library for all allowable flexible residues. Dihedral atom types and angles in the library were aligned to those in the existing backbone of the wildtype residue being mutated. Rotamers were eventually energy minimized in voxels for all conformations, including those participating in the global minimum energy

conformation (GMEC). Prior to the minDEE stage, a volume filter assessed the target volumes occupied by the wildtype residues and then pruned allowable mutant residues that were not within a volume window from this target volume. The minDEE stage examined the rotamer pairwise energies and pruned rotamers that were not within 5 kcal/mol of the GMEC energy. Prior to the A* stage, a steric filter pruned conformations having van der Waals (vdW) radii that overlapped by more than 1.5 Å, as relaxation from such a severe clash is not likely using energy minimization. The A* stage then enumerated conformations in order of lower bounds of energies so that this stage could be terminated once partial partition functions reach the desired accuracy of approximation (112). As the A* stage fed the K* stage, an ϵ -approximation guaranteed that the computed partial partition function, incorporating a subset of the remaining unpruned conformations, represented at least 97% of the full partition function. Each successive stage incurred greater computational cost, and thus was executed later in the pipeline. Pruning was achieved with provable guarantees while both adapting GMEC-based and applying ensemble-based criteria. For example, minDEE identified the bound or unbound GMEC from the ensemble of conformations having flexible side-chains (Supplementary Figure 4-2). Adapted to this ensemble-based protocol, minDEE pruned conformations of a given mutant sequence to yield its GMEC as well as a gap-free list of other low-energy

conformations that are within an energy window from this GMEC. Furthermore, when A^* and K^* were applied, the computed energies of bound and unbound conformations were used in corresponding partition functions to determine binding affinities.

Combinatorial computational scanning and redesign. Pruning was performed on all conformations based on wildtype residue packing volumes, so that over-packed or under-packed allowable residues in the mutation map were removed from subsequent computation. A target volume initially was computed through summation of volumes for a rotamer of each wildtype residue in the mutation map. A volume window that permitted a specified volume over-packing or under-packing from this target volume was applied for redesign, but not for scanning. Further pruning was performed on bound and unbound conformations using minDEE, sterics, and K^*/A^* . The minDEE algorithmic pruning eliminated dead-end rotamers not near the minimized GMEC using simple coupled Goldstein criterion (112,131) with pairs removal from the aforementioned energy bounds. After an initial minDEE pruning, divide and conquer splitting was performed (114). An energy window of 5.0 kcal/mol ensured that no conformations having energy within this window from the GMEC energy were pruned. The energy bounds had a cutoff energy threshold of 100.0 kcal/mol. Steric pruning was then applied to prune

Supplementary Figure 4-2: minimized side-chain Dead-End Elimination (minDEE) criterion for nucleic acid binding proteins



$$\underbrace{E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s)}_{\text{analogous to traditional DEE criterion, but with minimum bound}} - \underbrace{\sum_{j \neq i} \max_s E_{\emptyset}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\emptyset}(j_s, k_u)}_{\text{range accounts for energy changes that may occur during scanning or redesign energy minimizations}} > \underbrace{E_{\oplus}(i_t) + \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s)}_{\text{analogous to traditional DEE criterion, but with maximum bound}}$$

minDEE criterion

With conformations being energy minimized, the minDEE criterion guarantees that no rotamers belonging to the minimized GMEC are pruned. The rotamers of conformations are energy minimized in voxels (dotted conic regions) of conformational space within $\pm \theta^\circ$ from the each rotamer's dihedral angle. With such minimization, the maximum, minimum, and range of these energies can be computed. Building upon (112), these computations enable the scanning and redesign of protein residues near nucleic acid-protein interfaces, thereby extending the traditional DEE criterion to consider the effect of energy minimization to changes in conformational energies.

conformations in which atomic vdW radii clash by more than 1.5 Å, assuming that such extreme steric clashes cannot relax to the aforementioned hard steric threshold by steepest-descent-based energy minimization. Rotamer pairs were further pruned based on a steric energy threshold. Rotamers with a summed intra-rotamer and rotamer-to-template energy greater than 30.0 kcal/mol were pruned. Furthermore, rotamer pairs with interaction energy greater than 30.0 kcal/mol were also pruned. The K*/A* algorithms perform enumeration in order of lower bounds on energy by calculating whether each subsequent minimized conformation is expected to contribute to an ϵ -approximation of the total partition function (112). Since partition functions are an exponentiated sum of energies of the unpruned bound (or unbound) conformations, some conformations in the ensemble contribute more than others to the total partition function. For expediency, an ϵ value of 0.03 was chosen to guarantee that the partially computed partition function for the bound (and unbound) ensemble of conformations would approximate at least 97% of the completely computed partition function value for that ensemble. For efficiency, expensive energy minimization and downstream computations are performed only for each successive conformation that can contribute to this partially computed partition function. High performance computing resources were used to distribute and process tasks through an object-oriented Java wrapper (mpiJava) to a message passing interface (MPI) (182-184).

Conformational energy, binding energy, and binding affinity. Of the three means of scoring used in this study, conformational energy was computed using the aforementioned energy function for all bound and unbound conformations. Binding energy was computed by identifying the GMECs among these conformations and calculating the difference in their bound and unbound conformational energies. Binding affinity, termed K^* score, was computed as the quotient of partition functions encoding the conformational energy contributions of all unpruned conformations that contribute significantly to the bound and unbound partition functions. While the binding energy computation used only the energies of the bound and unbound GMEC with unitary weights, the binding affinity computation used the conformational energies of all significant unpruned bound and unbound conformations weighted by Boltzmann probabilities. Therefore, GMEC was used to represent an energetically favorable bound conformation for each mutant at a particular instant. In contrast, in order to account for the conformational flexibility and adaptation that protein side-chains can assume, the binding affinity simultaneously assessed all rotameric conformations *in silico* rather than just the rotameric GMEC.

Output structure models, overall binding energy and binding affinity computation. The overall energies were computed for significantly

contributing bound (and unbound) conformations in the ensemble using the aforementioned energy function. The binding energy was defined as:

$$\Delta\Delta E_{\text{binding}} = \Delta E_{\text{bound}} - \Delta E_{\text{unbound}}$$

The $\Delta E_{\text{unbound}}$ component is composed of changes to the unbound RE energy, since the RE is mutated while the DNA substrate, and its energy, remain constant. The K* algorithm computed a binding affinity approximation, termed K* score, to the association constant, K_a , using partition functions of ensembles of bound and unbound conformations, where the energy contribution of each conformation is weighted by a Boltzmann probability. The partition functions (pq), (p), and (q), over an ensemble of bound and unbound conformations were defined as:

$$(pq) = \sum_{i \in \text{bound RE:DNA}} e^{-\frac{E_{i,\text{bound RE:DNA}}}{RT}} ; \quad (p) = \sum_{i \in \text{unbound RE}} e^{-\frac{E_{i,\text{unbound RE}}}{RT}} ; \quad (q) = \sum_{i \in \text{unbound DNA}} e^{-\frac{E_{i,\text{unbound DNA}}}{RT}}$$

where E is the aforementioned computed energy in kcal/mol for each conformation. The universal gas constant R ($=8.314472 \text{ J K}^{-1} \text{ mol}^{-1}$) was used as a molar multiplicative of the Boltzmann constant typically used in the canonical ensemble. The absolute temperature T ($=298.15 \text{ K}$) was used as it is consistent with the implicit solvation energy term of the energy function and physiological temperature for the RE and DNA. A reasonable statistical mechanics approximation to the kinetic binding constant was thus defined as (185):

$$K_a = \frac{[RE - DNA]}{[RE][DNA]} \approx \frac{probability_{bound}}{probability_{unbound}} \approx \frac{(pq)}{(p)(q)} = K^*$$

Losses in binding energy and binding affinity were computed as the difference in binding energy between mutant and wildtype GMECs:

$$\Delta\Delta E_{\text{binding loss}} = \Delta\Delta E_{\text{binding, mutant}} - \Delta\Delta E_{\text{binding, wildtype}}$$

and quotient of entire mutant and wildtype ensembles of conformations:

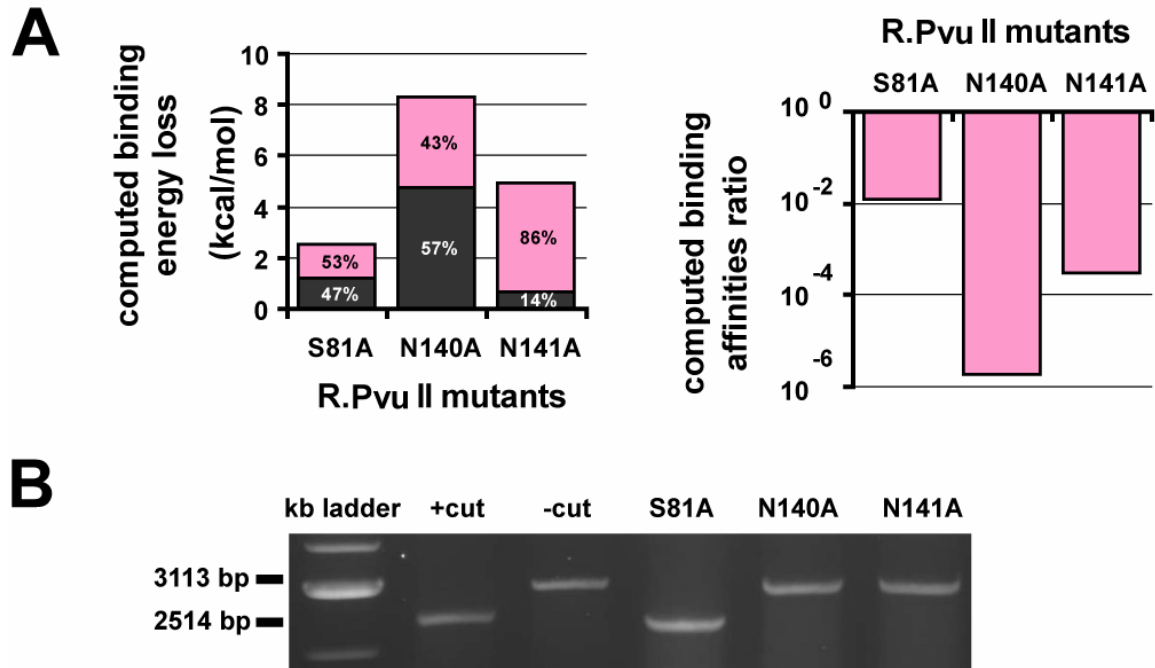
$$K^* \text{ score}_{\text{binding loss}} = K^* \text{ score}_{\text{binding, mutant}} / K^* \text{ score}_{\text{binding, wildtype}}$$

respectively. Atomic coordinates for the output models were generated for all unpruned conformations, and visualized for the GMECs.

4.4.2. Scanning for mutation tolerance and experimental validation

One application of this automated computational protocol was to scan amino acid residues for mutation tolerance. The three residues in R.PvuII, S81, N140, and N141, which had been previously identified to interact with the -3c::+3g BP in the cognate DNA substrate, , were mutated to all rotamers of themselves or alanine in order to assess losses due to this computational alanine scanning. Due to apparent asymmetry of the two monomers in the structural model, we performed both simultaneous and non-simultaneous

Figure 4-3: Computational and experimental alanine scans of selected residues in R.PvuII



(A) Computed homodimer binding energy loss (left, full columns) and binding affinity loss (right) as a measure of residue mutation tolerance. Monomer-specific contributions to binding energy losses are shown in stacked pink and gray columns (left). **(B)** Experimental alanine scan of the same residues validated the computational predictions. Mutant constructs were expressed in a cell-free system and tested for DNA digestion activity. Agarose gel electrophoresis of the digestion reactions indicated that S81A mutant maintained specific R.PvuII activity, while N140A and N141A mutants lost all enzymatic activity. The positive control (+cut) was cut with wildtype R.PvuII enzyme expressed in the same way as the mutants. The DNA substrate includes all 6-mer sequences.

alanine mutations for each of the three residues in the two monomers. As shown in Figure 4-3A, when scans were performed using simultaneous alanine mutations on both monomers, two wildtype-to-alanine mutations, N140A and N141A, resulted in binding energy losses of 8.32 and 4.98 kcal/mol and binding affinity losses of approximately 6 and 4 orders of magnitude from the wildtype, respectively. When compared to the single position computational scanning performed for all other positions on both monomers simultaneously, these binding energy losses represented approximately 1 standard deviation (5.28 kcal/mol) or greater than the mean binding energy loss (0.72 kcal/mol) of all the residues except prolines or G56 in the model. The S81A mutation yielded a binding energy loss of 2.53 kcal/mol, which was less than 1 standard deviation from the mean value, and a binding affinity loss of approximately 2 orders of magnitude. The computational alanine scan results suggested that mutations to S81 would be tolerated whereas mutations to N140 and N141 would lose too much binding energy and binding affinity. When scans of non-simultaneous mutations were performed, the results revealed that corresponding positions on each monomer contributed unequally to the loss of binding energy, thus highlighting the asymmetry found in the models for symmetric biological units (Figure 4-3A).

Supplementary Table 4-1: Pruning efficiency of redesigned R.PvuII S81 mutants for cognate and non-cognate DNA substrates

	pruning stage	sequences remaining	sequence factor	pruning	conformations remaining	conformation factor	pruning	
		mutable:	S81					
		flexible:	S81					
bound conformations	initial	19 ; 19 ;	–		152 ; 152 ;	–		
	allowable	19 ; 19			152 ; 152			
	packing	8 ; 8 ;	2.375 (42%) ; 2.375 (42%) ;	26 ; 26 ;	5.846 (83%) ; 5.846 (83%) ;			
	volume	8 ; 8	2.375 (42%) ; 2.375 (42%)	26 ; 26 ;	5.846 (83%) ; 5.846 (83%)			
	minDEE	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	12 ; 12 ;	2.167 (54%) ; 2.167 (54%) ;			
	energy	8 ; 8	0.000 (0%) ; 0.000 (0%)	8 ; 11	3.250 (69%) ; 2.364 (58%)			
bound conformations	A* steric	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	12 ; 12 ;	0.000 (0%) ; 0.000 (0%) ;			
	constraints	8 ; 8	0.000 (0%) ; 0.000 (0%)	8 ; 11	0.000 (0%) ; 0.000 (0%)			
	K* partition	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	9 ; 9 ;	1.333 (25%) ; 1.333 (25%) ;			
	fcn. contrib.	8 ; 8	0.000 (0%) ; 0.000 (0%)	8 ; 10	0.000 (0%) ; 1.100 (9%)			
	initial	19 ; 19 ;	–		152 ; 152 ;	–		
	allowable	19 ; 19			152 ; 152			
unbound conformations	packing	8 ; 8 ;	2.375 (42%) ; 2.375 (42%) ;	26 ; 26 ;	5.846 (83%) ; 5.846 (83%) ;			
	volume	8 ; 8	2.375 (42%) ; 2.375 (42%)	26 ; 26 ;	5.846 (83%) ; 5.846 (83%)			
	minDEE	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	15 ; 15 ;	1.733 (42%) ; 1.733 (42%) ;			
	energy	8 ; 8	0.000 (0%) ; 0.000 (0%)	15 ; 15	1.733 (42%) ; 1.733 (42%)			
	A* steric	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	15 ; 15 ;	0.000 (0%) ; 0.000 (0%) ;			
	constraints	8 ; 8	0.000 (0%) ; 0.000 (0%)	15 ; 15	0.000 (0%) ; 0.000 (0%)			
unbound conformations	K* partition	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	11 ; 11 ;	1.364 (27%) ; 1.364 (27%) ;			
	fcn. contrib.	8 ; 8	0.000 (0%) ; 0.000 (0%)	11 ; 11	1.364 (27%) ; 1.364 (27%)			
			mutable:	S81				
			flexible:	K70, F80, S81, T82, N141, K143				
	bound conformations	initial	19 ; 19 ;	–		9.3x10 ⁶ ; 9.3x10 ⁶	–	
		allowable	19 ; 19			9.3x10 ⁶ ; 9.3x10 ⁶		
packing		8 ; 8 ;	2.375 (42%) ; 2.375 (42%) ;	1.6x10 ⁶ ; 1.6x10 ⁶ ;	5.846 (83%) ; 5.846 (83%) ;			
volume		8 ; 8	2.375 (42%) ; 2.375 (42%)	1.6x10 ⁶ ; 1.6x10 ⁶	5.846 (83%) ; 5.846 (83%)			
minDEE		8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	2.6x10 ⁵ ; 2.6x10 ⁵ ;	6.143 (84%) ; 6.143 (84%) ;			
energy		8 ; 8	0.000 (0%) ; 0.000 (0%)	3.3x10 ⁵ ; 2.6x10 ⁵	4.774 (79%) ; 6.143 (84%)			
bound conformations	A* steric	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	2.3x10 ⁵ ; 2.3x10 ⁵ ;	1.126 (11%) ; 1.130 (12%) ;			
	constraints	8 ; 8	0.000 (0%) ; 0.000 (0%)	3.1x10 ⁵ ; 2.4x10 ⁵	1.093 (9%) ; 1.099 (9%)			
	K* partition	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	1.8x10 ⁴ ; 2.3x10 ⁴ ;	12.764 (92%) ; 9.981 (90%) ;			
	fcn. contrib.	8 ; 8	0.000 (0%) ; 0.000 (0%)	8.0x10 ³ ; 1.1x10 ⁴	38.064 (97%) ; 21.938 (95%)			
	initial	19 ; 19 ;	–		9.3x10 ⁶ ; 9.3x10 ⁶	–		
	allowable	19 ; 19			9.3x10 ⁶ ; 9.3x10 ⁶			
unbound conformations	packing	8 ; 8 ;	2.375 (42%) ; 2.375 (42%) ;	1.6x10 ⁶ ; 1.6x10 ⁶ ;	5.846 (83%) ; 5.846 (83%) ;			
	volume	8 ; 8	2.375 (42%) ; 2.375 (42%)	1.6x10 ⁶ ; 1.6x10 ⁶	5.846 (83%) ; 5.846 (83%)			
	minDEE	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	5.1x10 ⁵ ; 5.1x10 ⁵ ;	3.150 (68%) ; 3.150 (68%) ;			
	energy	8 ; 8	0.000 (0%) ; 0.000 (0%)	5.1x10 ⁵ ; 5.1x10 ⁵	3.150 (68%) ; 3.150 (68%)			
	A* steric	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	4.4x10 ⁵ ; 4.4x10 ⁵ ;	1.139 (12%) ; 1.139 (12%) ;			
	constraints	8 ; 8	0.000 (0%) ; 0.000 (0%)	4.4x10 ⁵ ; 4.4x10 ⁵	1.139 (12%) ; 1.139 (12%)			
unbound conformations	K* partition	8 ; 8 ;	0.000 (0%) ; 0.000 (0%) ;	9.9x10 ⁴ ; 9.9x10 ⁴ ;	4.494 (78%) ; 4.494 (78%) ;			
	fcn. contrib.	8 ; 8	0.000 (0%) ; 0.000 (0%)	9.9x10 ⁴ ; 9.9x10 ⁴	4.494 (78%) ; 4.494 (78%)			

reported for DNA substrates: 5'-cagctg-3' ; 5'-tagcta-3' ; 5'-aagctt-3' ; 5'-gagctc-3'

The pruning factor represents the ratio of the number of protein sequences or conformations present before and after the given pruning stage. The pruning-% (in parentheses) represents the percentage of remaining protein sequences or conformations eliminated by the given pruning stage. Reported for DNA substrates 5'-cagctg-3' ; 5'-tagcta-3' ; 5'-aagctt-3' ; 5'-gagctc-3'.

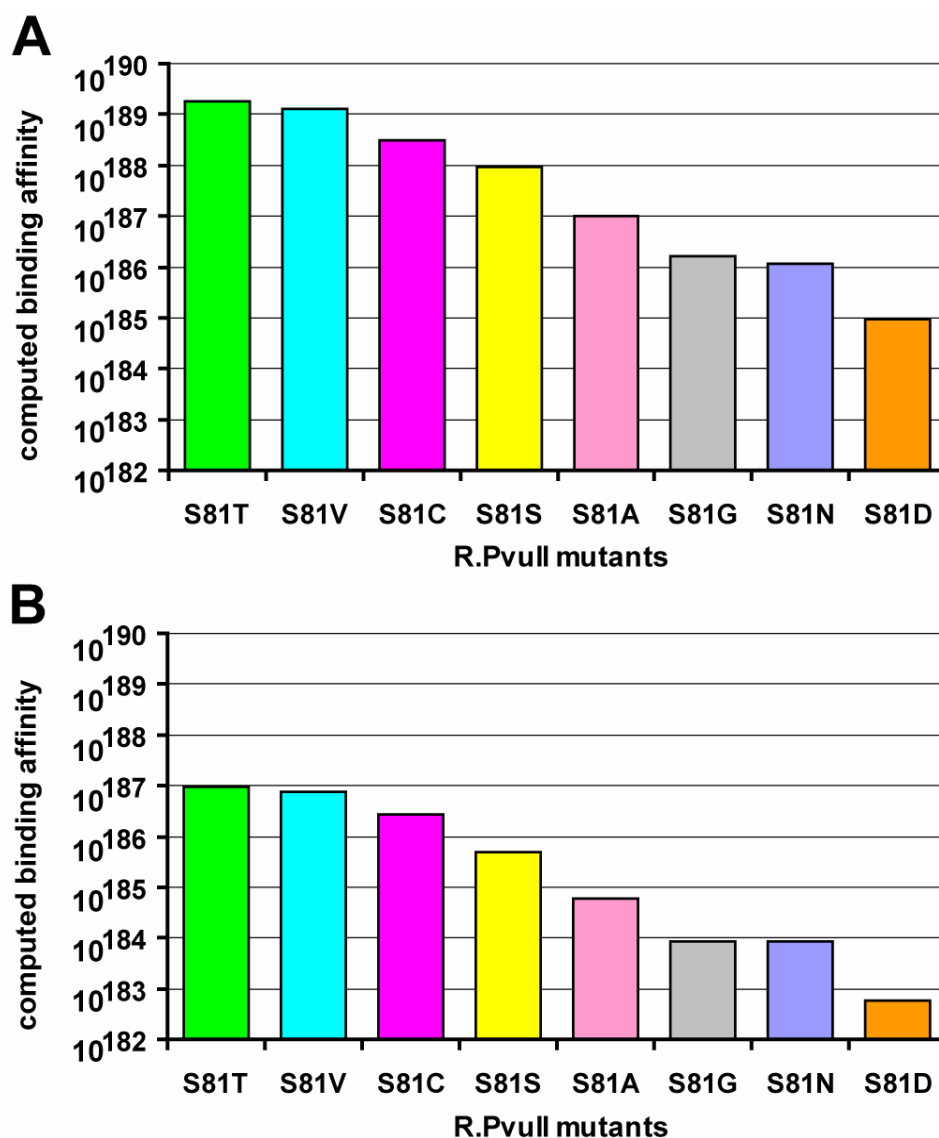
The stages pruned varying numbers of bound and unbound conformations. Fewer unbound as opposed to bound conformations were pruned. This is not unexpected as the absence of the DNA substrate makes more space available for rotameric side-chains from the protein to be placed without steric clashes.

Experimental alanine scans agreed with the assessment of tolerance by computational scans on S81, N140, and N141, as indicated by the presence or absence of the +cut DNA band, representing complete digestion of all three cognate sequences found in the substrate (Figure 4-3B). In addition to the qualitative importance of binding energy loss, the quantitative differences highlighted the effect of REase residue position and proximity to the DNA substrate. For example, atoms of N140 and N141 are closer to the DNA and lose more binding energy than S81 when scanned. Despite this greater distance, it has been previously reported that this S81 residue is also critical to binding and cleavage, and that redesign attempts such as S81L were not successful (186). Since our scanning results indicated that S81 was mutation tolerant, S81 was focused on as the target for redesign.

4.4.3. Computational redesign for functional S81 mutants and experimental validation

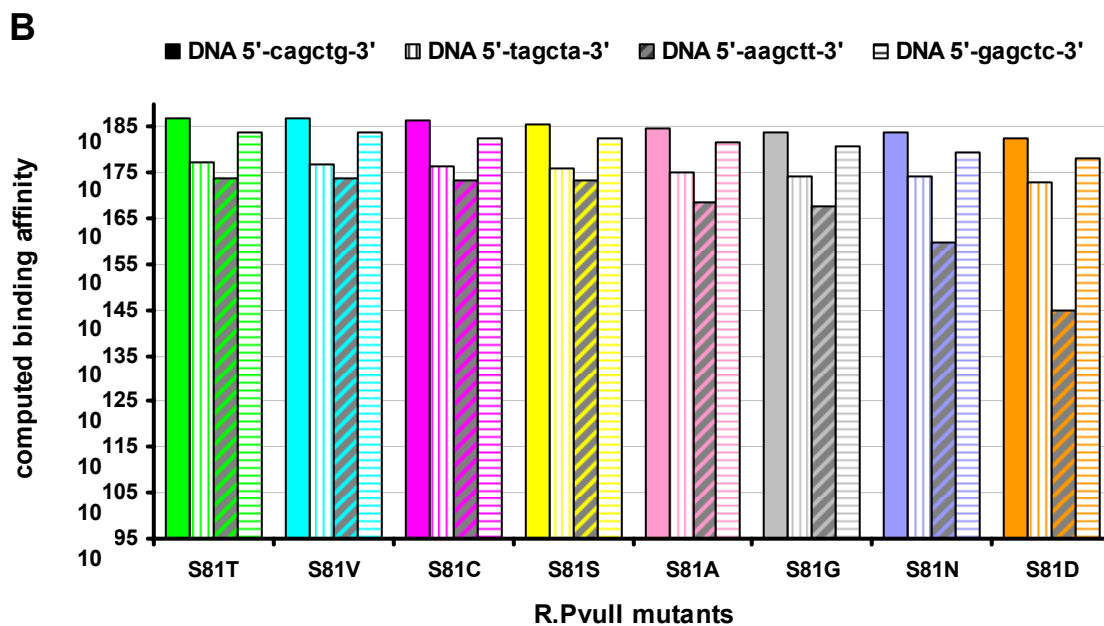
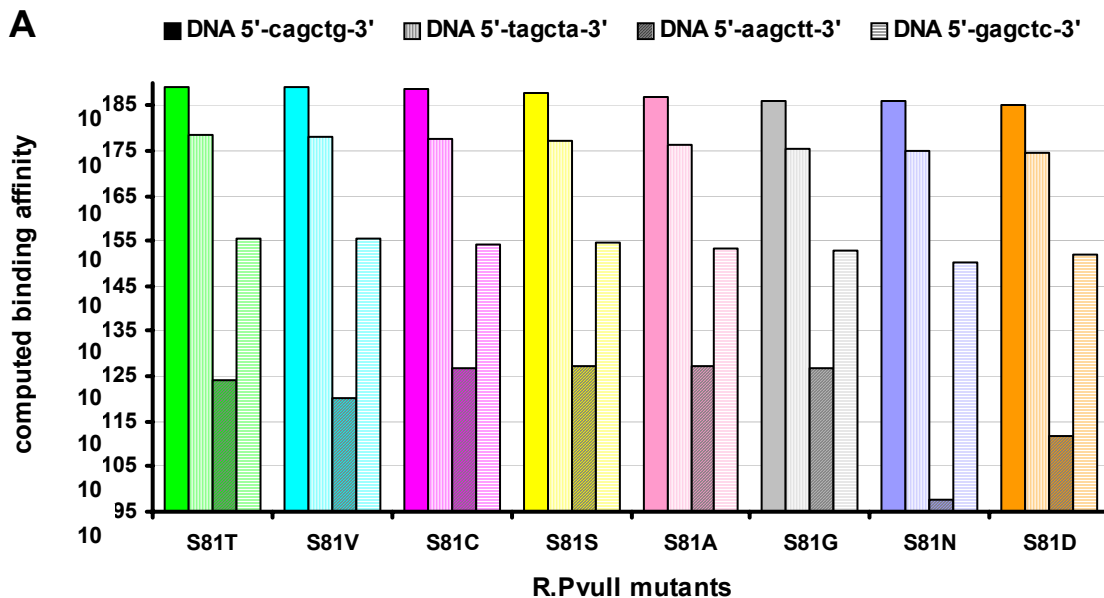
Another application of this automated protocol was to redesign mutation tolerant residues, such as S81 identified previously, by reducing the mutation search space at S81 and flexible search space of all residues 3.0 Å from S81 from 19 possible protein sequences containing 9,307,872 possible bound (and an equal number of unbound) rotameric conformations to 8 mutants containing 18,037 and 98,727 possible bound and unbound

Figure 4-4: Computed binding affinities of redesigned R.PvuII S81 mutants to cognate DNA substrate



(A) With residue S81 permitted to be mutable to all amino acids except proline and to be flexible, a 1-point simultaneous mutation search was performed over 304 initial allowable total bound and unbound conformations. **(B)** With residue S81 permitted to be mutable to all amino acids except proline and residues having at least one atom within 3 Å of an atom of residue S81, i.e. K70, F80, S81, T82, N141, K143, permitted to be flexible, a 6-point simultaneous mutation search was performed over 18,860,688 initial allowable total bound and unbound conformations.

Supplementary Figure 4-3: Redesigned R.PvuII S81 mutants computed binding affinities to the cognate and non-cognate DNA substrates



(A) With residue S81 permitted to be mutable to all amino acids except proline and to be flexible, a 1-point simultaneous mutation search was performed over 1,216 initial allowable total bound and unbound conformations for the cognate, 5'-cagctg-3', and for the three possible non-cognate sequences at the -3::+3 BP position. Comparative evaluation

revealed that all these mutants have a general, but varying, preference for the cognate sequence. This preference was experimentally tested with a DNA substrate containing all 64 palindromic 6-mer sequences, including these aforementioned for sequences, and validated to be the case. Of the three non-cognate sequences, there existed a general, but varying, preference for 5'-tagcta-3' which maintains the heterocyclic pyrimidine::purine ring configuration at the -3::+3 BP position with respect to the cognate sequence.

(B) With residue S81 permitted to be mutable to all amino acids except proline and residues having at least one atom within 3 Å of an atom of residue S81, i.e. K70, F80, S81, T82, N141, K143, permitted to be flexible, a 6-point simultaneous mutation search was performed over 74,462,976 initial allowable total bound and unbound conformations for the cognate, 5'-cagctg-3', and for the three possible non-cognate sequences at the -3::+3 BP position. Comparative evaluation once again revealed the preference for the cognate sequence remains while greater flexibility around the redesigned residue suggested the possibility of non-specific “star” activity exhibited by REases under more destabilizing conditions.

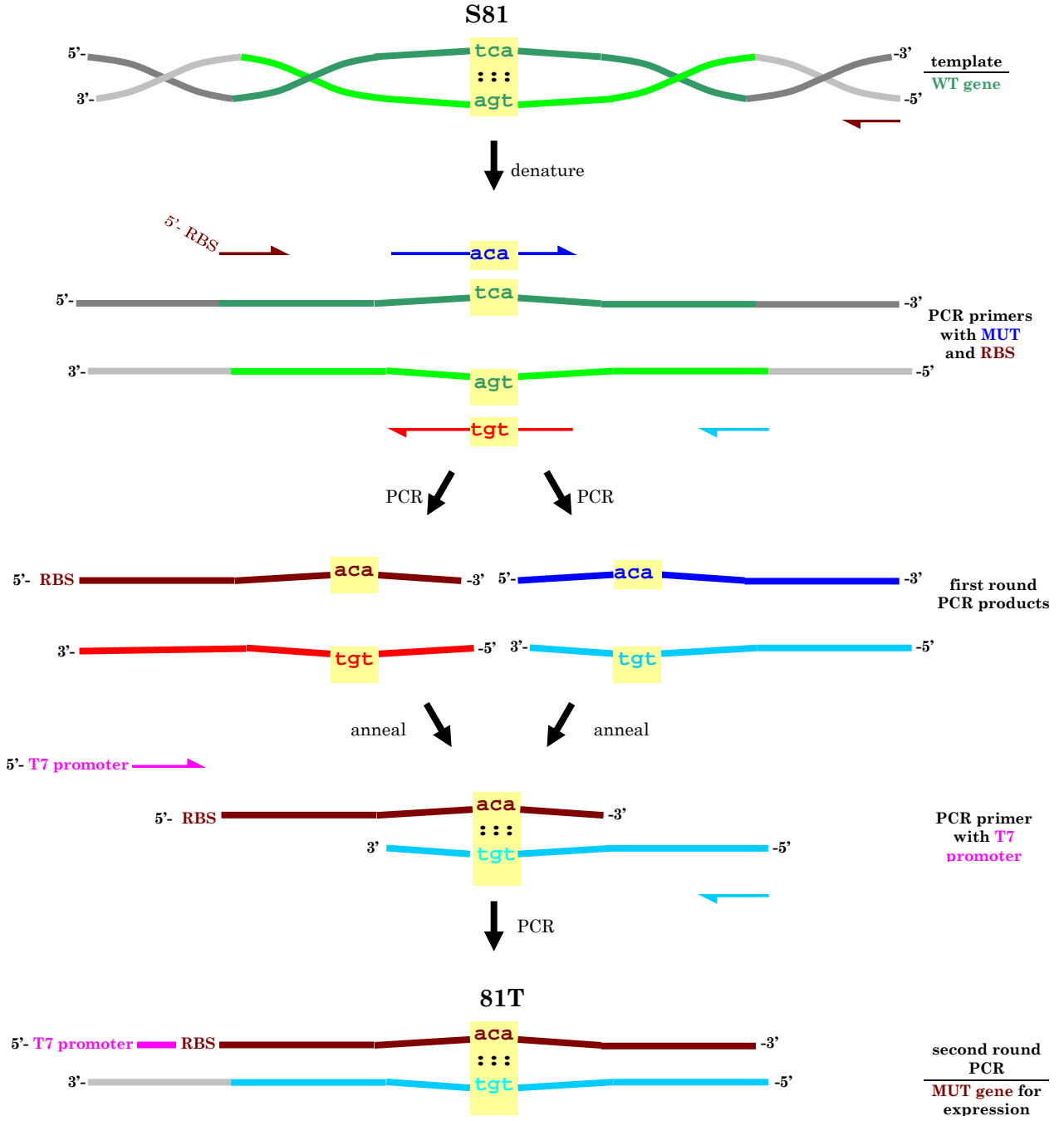
conformations, respectively (Supplementary Table 4-1). In addition to addressing this combinatorial complexity through the consideration of more flexible residues so that the 9,430,344 possible bound (and an equal number of unbound) rotameric conformations were initially considered for each DNA substrate cognate or non-cognate sequence, various stages of the protocol enabled predictions to be made that have biological accuracy. From the 19 initial allowable residues, the volume filter enforced a 20% volume window that was satisfied by 8 of these residues. Then, minDEE eliminated 1,332,936 out of 1,592,136 (or 84%) bound and 1,086,696 out of 1,592,136 (or 68%) unbound rotamers as not being at or within 5.0 kcal/mol of the GMEC. While no remaining conformations after this stage presented a steric clash severe enough to be pruned prior to A*, the sum of exponentiated energies of only 18,037 of the remaining 230,228 bound and 98,727 of the 443,720 unbound conformations for 8 mutants contributed to 97% or more to the corresponding full bound and unbound partition functions, and thus were scored by K*. The sums of exponentiated energies for the other conformations, being of higher energy, were proven not to contribute enough to this 97% of the corresponding full partition functions to warrant inclusion. Interestingly, comparative evaluation for all three other possible non-cognate BPs at the -3::+3 position on the DNA substrate revealed that all eight

redesigned mutants show higher affinity for the cognate substrate (Supplementary Figure 4-3).

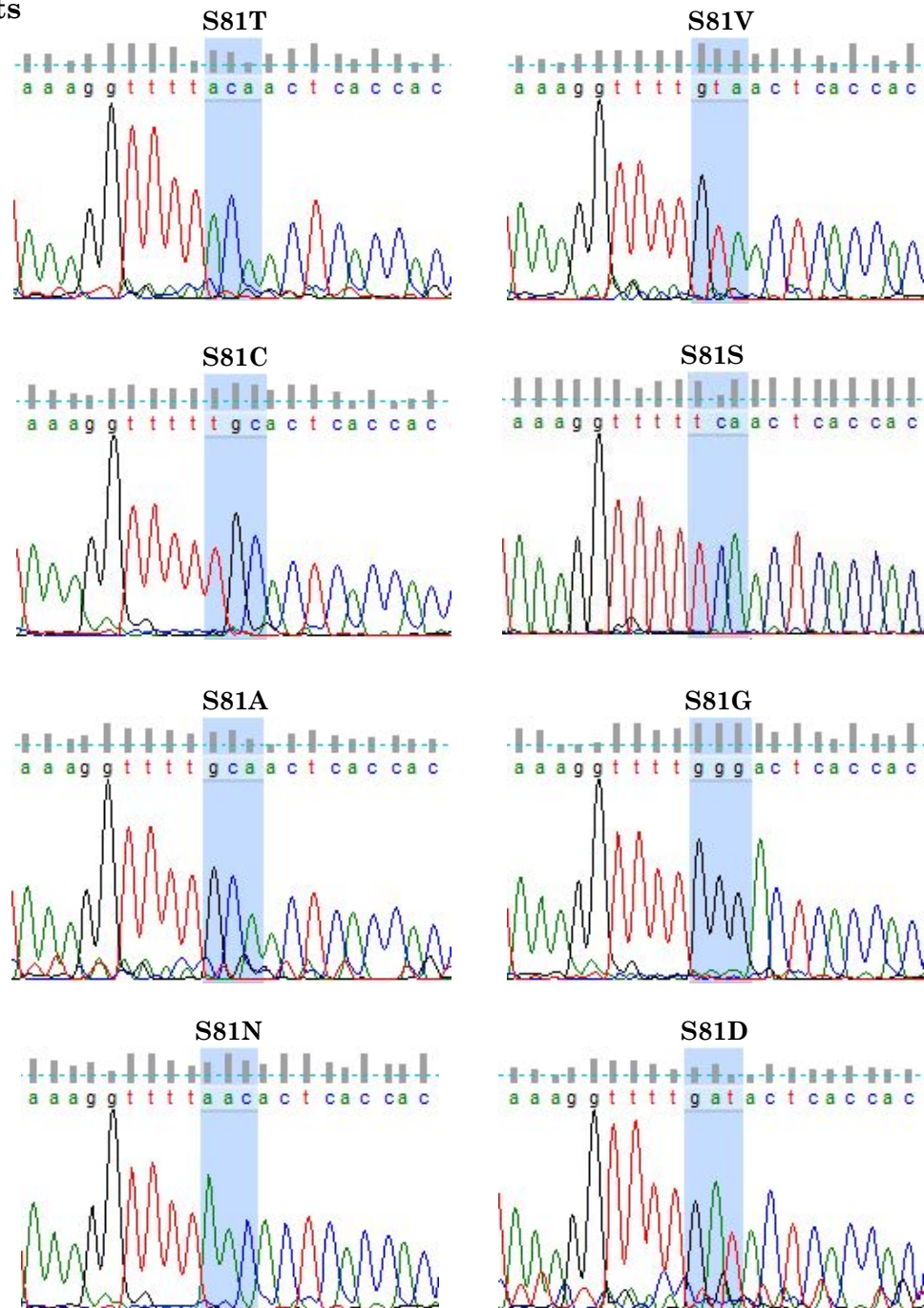
Accurately predicting functional mutants out of 19 sequences or 152 bound conformations is not trivial. The protocol generated 8 redesigns, S81T, S81V, S81C, S81S, S81A, S81G, S81N, and S81D, in order of their binding affinities to the cognate DNA substrate (Figure 4-4). Among the 8 redesigns, S81S was a true positive, as it was the rotameric native recovery of the functional wildtype sequence and structure, with a structural root mean square deviation averaged across both monomers and normalized by number of atoms in this residue of only 0.060 Å. Another mutant that was pruned rather than predicted was an experimentally proven true negative, as it was the same S81L mutant that was already found to not function in earlier studies (186). The function of the remaining 7 mutants had not been tested previously and thus warranted further investigation.

The 8 redesigns were experimentally evaluated for restriction specificity and efficiency using DNA restriction digestion assays. The R.PvuII gene, *pvuII*R, and the redesigned mutants were synthesized and mutated *de novo* (Supplementary Figure 4-4) and verified by sequencing (Supplementary Figure 4-5). To avoid potential cytotoxicity problems, these REase mutants were expressed from synthetic genes using the PURE cell-

Supplementary Figure 4-4: Enzyme synthesis and mutagenesis



Supplementary Figure 4-5: DNA sequencing chromatograms validates gene sequence of scanned and redesigned R.PvuII S81 mutants

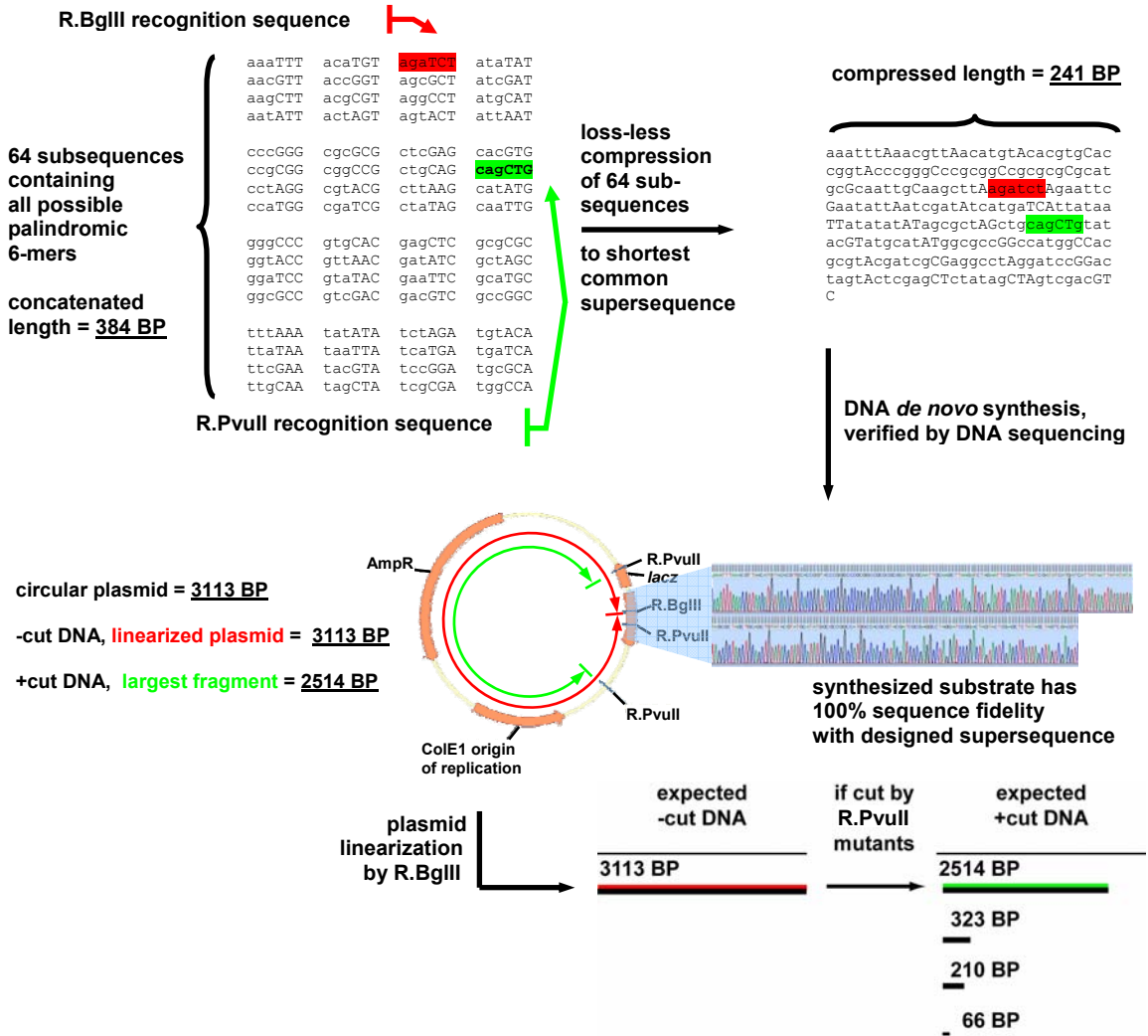


The most frequently used codon in the synthesized *PvuII*R wildtype gene sequence encoding each residue type was determined. These codons (highlighted in blue) were then specifically implemented during mutagenesis to the gene sequences for the scanned and redesigned mutants.

free protein synthesis system (187-191). A typical reaction using the PURE system yielded about 50 pM of each full-length protein product with a C-terminal His6-tag.

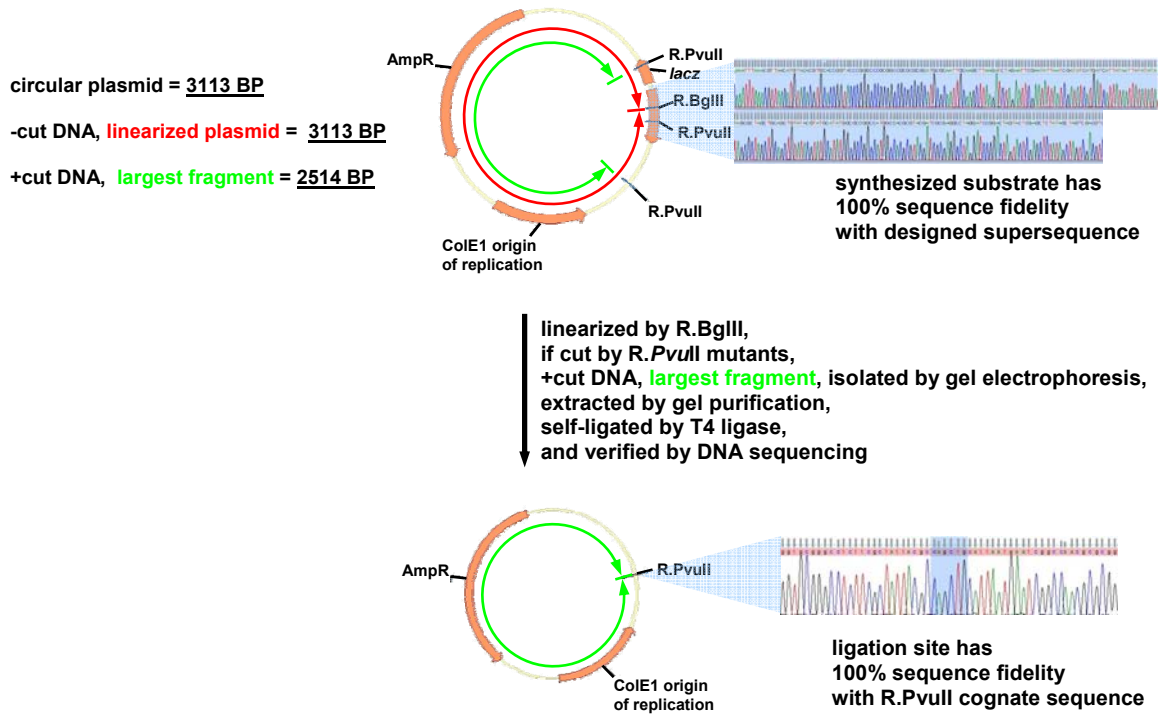
For the restriction digestion assays, a synthetic DNA substrate containing the cognate sequence (5'-cagctg-3') as well as all 63 other possible palindromic 6-mer non-cognate sequences was constructed to fairly assess the mutants' functional specificity (Supplementary Figure 4-6). Digestion with R.PvuII or a true functional mutant was expected to produce a distinct signature of cut DNA fragments (+cut) which is clearly discernable from the uncut (-cut) or those cut at different sites. The exact cutting site was identified by sequencing of re-ligation products if necessary (Supplementary Figure 4-7). As shown in Figure 4-5, the results of the restriction digestion assays supported the redesign predictions from the computational protocol. The apparent molecular weights and relative concentrations of the His6-tagged protein products of the mutant genes were verified by Western blotting with anti-His6 antibody (Figure 4-5A). With equivalent amount of each mutant protein present in the restriction digestion assays, it was found that the top 6 mutants with high scores in predicted ensemble-based binding energies, S81T, S81V, S81C, S81S, S81A, and S81G, all showed restriction digestion activities; the 2 lowest-scoring mutants, S81N and S81D, exhibited non-detectable activities over a time-course of 2 hours at 37°C (Figure 4-5B).

Supplementary Figure 4-6: Design and evaluation of DNA substrate for REases



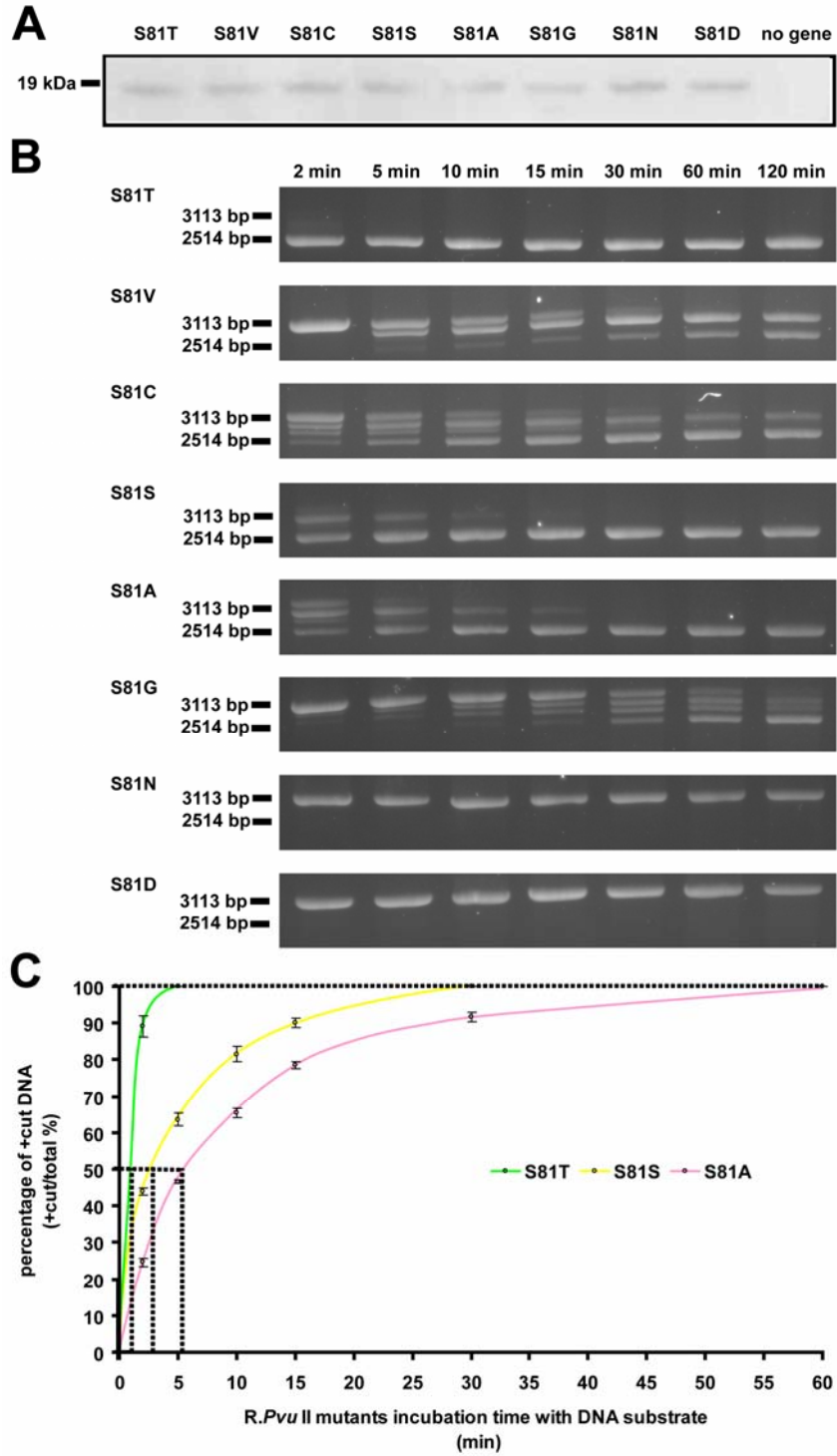
A synthetic DNA substrate sequence was designed that contained all 64 possible palindromic 6-mers as subsequences. For accuracy and efficiency of synthesis, this concatenated sequence of length 384 BP was computationally compressed in a loss-less manner to the shortest common supersequence of length 241 BP containing all subsequences in overlapping arrangements. This DNA supersequence was *de novo* synthesized and verified by sequencing. Upon inserting this synthesized supersequence into a circular plasmid backbone of length 2872 BP, the total length of the plasmid for the assays was 3113 BP. With the restriction map determined, this circular plasmid was linearized (red line) using *R.BglII*, since its restriction site occurred only once. This linearized plasmid, when fully digested at the *R.PvuII* cognate sequence by scanned or redesigned mutants, are expected to produce DNA fragments of lengths 2514 (green line), 323, 210, and 66 BPs. The largest of these +cut DNA fragments are readily comparable to the -cut DNA linearized plasmid during gel electrophoresis.

Supplementary Figure 4-7: DNA sequencing chromatograms validates restriction sequence of +cut DNA, i.e. largest fragment, from DNA substrate after complete digestion by redesigned R.PvuII S81 mutants



The largest fragment was isolated by gel electrophoresis and extracted by gel purification. The purified fragment was then incubated with T4 ligase for self-ligation and sequenced using the method and tools detailed elsewhere in this work.

Figure 4-5: Enzymatic activities of redesigned R.PvuII S81 mutants



(A) Western blot validating that equivalent concentrations of R.PvuII mutant proteins were synthesized and used in restriction digestion assays.

(B) Agarose gel electrophoresis of equal aliquots of restriction digestion reactions of the eight mutants taken at specified time points. The top band of 3,113-BP represents the full-length DNA substrate. The 2,514-BP band represents the longest fully-digested product; other smaller digestion products were out of the exposure area. The bands in between are incomplete digestion products as the linear DNA substrate contains 3 R.PvuII restrictions sites.

(C) Kinetics of restriction digestions of the three most active enzymes, S81T, S81S, and S81A (n = 3 for each mutant). The relative quantity of each band in the agarose gel was quantified by densitometry analysis. The percentage of +cut DNA over total DNA was calculated based on densitometry results. The kinetics data was plotted and reaction half-times were measured on the plot.

The patterns of the digestion intermediates and final products all indicated that the six functional mutants all recognized the cognate R.PvuII site. The level of enzymatic activities varied among the six functional mutants. For the three constructs that completely digested the DNA substrate within 2 hours at 37°C, S81T, S81S, and S81A, the half-times, or times taken to turnover half of the DNA substrate to +cut DNA, were 1.0, 2.6, and 5.7 minutes, respectively (Figure 4-5C). It is apparent that the top-scoring mutant, S81T, exhibited significantly higher activity than the wildtype while maintaining specificity (Supplementary Figure 4-7). This further demonstrated that ensemble-based binding affinities, rather than GMEC-based bound energies, were better predictors of these experimental outcomes (Table 1).

Output 3D models. To gain intuition on the interactions taking place in the redesigned mutants, a final step of the protocol built 3D output models for analysis. Atomic coordinates for output models were built for bound and unbound conformations post-pruning and post-scoring. For uniformity, the output models were built using the same rotamer library and energy minimization within voxels that were applied upstream in the protocol to mutate the input models. Similarly, the same composite energy function was applied to evaluate the energy of each output model in order to rank by energy and identify the bound and unbound GMEC for each mutant. The bound GMEC for each mutant was then post-hoc analyzed for biophysical

properties, such as steric packing and biochemical complementarity, which may confer favorable interactions with the DNA. These models revealed that the mutants with predicted affinities higher than the wildtype exhibited packing similar to the wildtype, while those with predicted affinities lower than the wildtype were under- or over-packed with respect to the wildtype (Figure 4-6).

4.5. Discussion

This study developed a computational molecular engineering protocol that successfully redesigned a challenging Type II restriction endonuclease, R.PvuII, after scanning its residues *in silico* for mutation tolerance. The protocol recovered the native sequence and structure and predicted high binding affinity mutants using partition functions over conformational ensembles. These benefits were reaped from its groundings in statistical mechanics, though at substantial computational cost. The protocol is general and can be applied to any nucleic acid binding protein (NABP) for which structure coordinates are available for the bound conformation. This can not only contribute to the understanding of existing protein-nucleic acids interactions, but may enable the modeling and engineering of synthetic

Table 4-1: Computed ensemble-based binding affinities, rather than global minimum energy conformation-based bound energies, are better predictors of experimental outcomes

S81 mutant	binding affinity rank (K* score)	bound energy rank (kcal/mol)	function presence	function half-time
	mutable:	S81		
	flexible:	S81		
S81T	1 st (1.779×10^{189})	2 nd (-403.682)	+cut	1.0 min
S81V	2 nd (1.308×10^{189})	4 th (-402.076)	+cut	ND
S81C	3 rd (3.152×10^{188})	5 th (-401.389)	+cut	ND
S81S	4 th (9.462×10^{187})	6 th (-400.863)	+cut	2.6 min
S81A	5 th (9.922×10^{186})	7 th (-397.230)	+cut	5.7 min
S81G	6 th (1.637×10^{186})	8 th (-395.396)	+cut	ND
S81N	7 th (1.155×10^{186})	1 st (-405.865)	-cut	ND
S81D	8 th (9.259×10^{184})	3 rd (-402.816)	-cut	ND
	mutable:	S81		
	flexible:	K70, F80, S81, T82, N141, K143		
S81T	1 st (9.252×10^{186})	2 nd (-401.855)	+cut	1.0 min
S81V	2 nd (7.408×10^{186})	4 th (-400.102)	+cut	ND
S81C	3 rd (2.670×10^{186})	5 th (-399.384)	+cut	ND
S81S	4 th (4.894×10^{185})	6 th (-398.977)	+cut	2.6 min
S81A	5 th (5.937×10^{184})	7 th (-395.249)	+cut	5.7 min
S81G	6 th (8.656×10^{183})	8 th (-393.372)	+cut	ND
S81N	7 th (8.576×10^{183})	1 st (-403.525)	-cut	ND
S81D	8 th (5.845×10^{182})	3 rd (-400.723)	-cut	ND
	reported for DNA substrate:	5'-cagctg-3'		

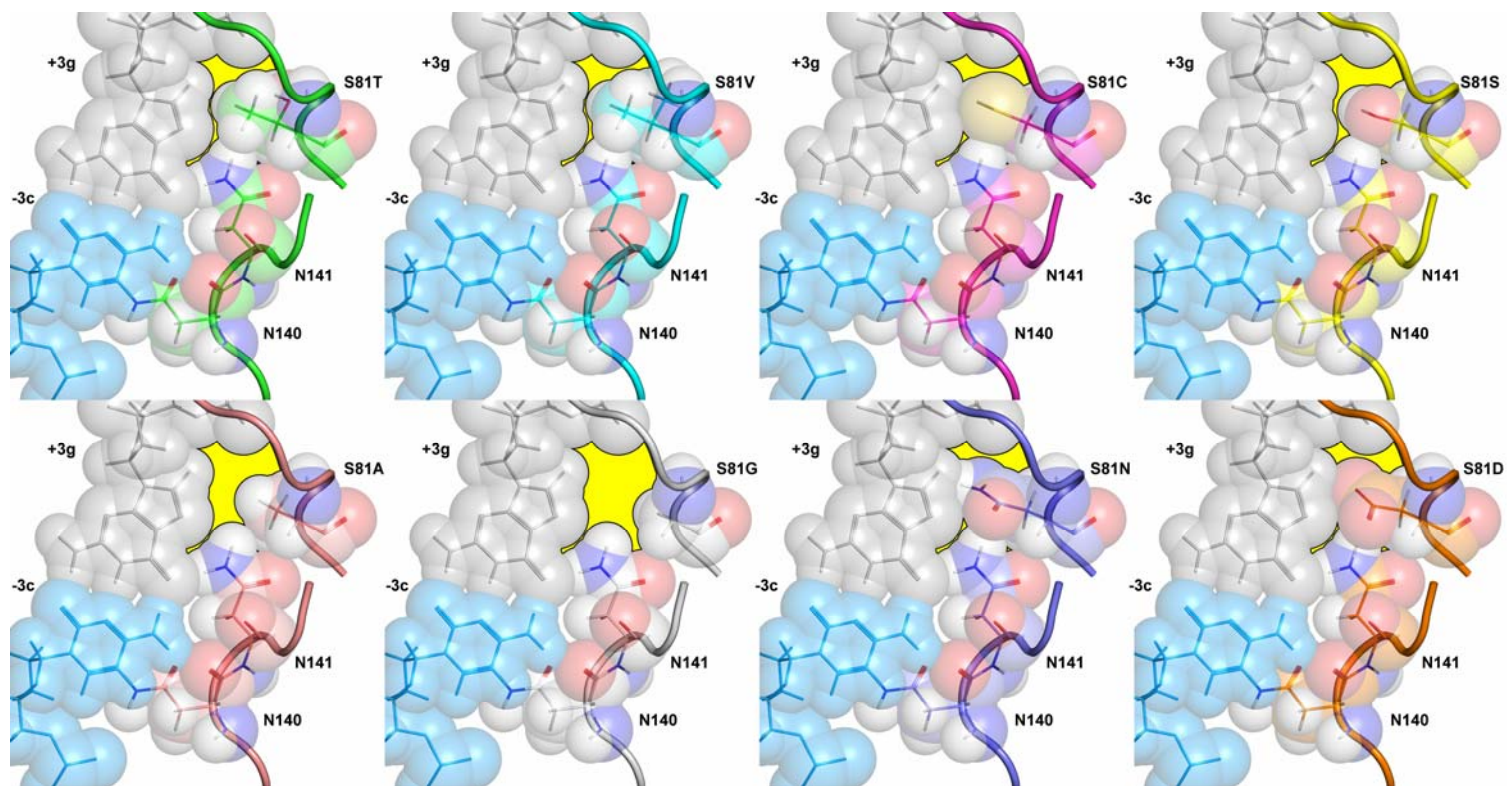
ND = not determined.
n = 3

interactions as well (177).

The example used here represents a particularly significant computational scanning and redesign challenge. Since R.PvuII is the smallest known REase, and is a homodimer encoded by a single gene, small mutations in the gene sequence can have dramatic impact on the protein structure and function. Furthermore, given the 2.60 Å resolution and completeness at this resolution (34.8 %) of the input structure model, it is impressive that our approach was able to predict mutations that were experimentally found to be accurate. The challenge involves making mutations that were both tolerable, as determined by scanning, and functional, as discovered by redesign.

Alanine scanning is useful in assessing the importance of interacting residues between biomolecules. Computational alanine scanning of NABPs using our protocol is a facile means of identifying mutation tolerant and intolerant residues. This approach is rapid and cost-effective, taking on the order of minutes to hours to compute binding properties for a given scanned residue on a 2.66 GHz node of a parallel high performance computing cluster. This approach, to an extent, also accurately models reality, as asymmetric contributions of symmetric binding partners mentioned elsewhere in the literature (178) were computationally revealed during our scanning of the

Figure 4-6: Modeling of global minimum energy conformations near the -3c::+3g BP of cognate DNA substrate 5'-cagctg-3' illustrates packing of redesigned R.PvuII S81 mutants



134

Packing (enclosed yellow area) is one of the distinguishing features of redesigned R.PvuII S81 mutants.

homodimeric REase with the palindromic DNA. In contrast to other computational methods with similar features, this approach is not heuristic and does guarantee a gap-free list of mutants for experimental consideration. Similar scanning has been performed for protein-protein interfaces of single-conformation and molecular dynamics approximated multiple-conformations (192-195). However, to our knowledge this is the first report of computational alanine scanning on protein-nucleic acid interfaces using molecular ensembles.

Redesign of NABPs is vital to the understanding and engineering protein-nucleic acid interactions. However, this field has been fraught with obstacles as there is no known or well characterized correspondence, or recognition codes between NABP amino acids and substrate nucleic acids (5). Our structure-based protocol obviates the reliance upon such correspondence codes by evaluating the interactions of each NABP specifically on a case-by-case basis and at an all-atom level.

The field has also attempted to redesign single conformations, such as the GMEC, without much success (186). The results presented here demonstrated that binding affinities computed from ensemble-based computations, rather than bound energies calculated from single-conformation assessments, are better predictors of experimental outcomes, and thus may describe the underlying dynamics better than these previous

approaches (Table 4-1). Of the surviving eight mutants predicted to bind with high affinity, the top ranked six were validated to function, with one discovered to out-perform the wildtype. This is the first report of an ensemble-based approach to redesigning NABPs based on statistical mechanics, where a collection of bound and unbound conformations are used in ensembles for Boltzmann-distributed partition functions, and the partition functions and computed conformational energies are utilized to compute provably-good approximations to binding affinities.

Though the redesigned GMEC structures represent only a single conformation among possibly many others in the ensembles that populate these partition functions, they are worth inspecting for intuition on molecular packing between REase and DNA. Nevertheless, the effects of such packing were evaluated by the energy function for all conformations and not just the GMEC. The S81S mutant, which the protocol outputs as the rotameric model equivalent of the wildtype structure, is recovered with the fourth highest K^* score for binding affinity. The three mutants having higher K^* scores, S81T, S81V, and S81C, are packed similarly to the wildtype, but present REase side-chain moieties to the DNA that are more favorably assessed by protocol. In particular, a reason why the S81T mutant out-performs the wildtype may be due to its ability to present both a hydroxyl group that can participate in hydrogen bonding and a methyl group for hydrophobic complementarity.

Two other mutants, S81A and S81G, having K^* lower than S81S, are apparently under-packed as compared to the wildtype. The remaining two mutants with the lowest K^* scores, S81N and S81D, appear over-packed as compared to S81S. According to the experimental results, it seems that, while the under-packing still permitted some attenuated R.PvuII enzymatic activity, over-packing significantly reduced or abolished it. The precise effects of under or over packing on the overall structure of the protein remain to be investigated, but the utility of packing in design is in agreement with findings elsewhere (196).

5. Open-source molecular engineering

*There is something fascinating about science.
One gets such wholesale returns of conjecture
out of such a trifling investment of fact.*

– Mark Twain
American author

This chapter has been adapted partially from a manuscript that was joint work with Ivelin Georgiev, Jingdong Tian and Bruce R. Donald:

Reza F., Georgiev I., Tian J., Donald B. R. Open-source computational redesign of nucleic acid binding proteins. *To be submitted.*

and partially from a research meeting abstract that was joint work with Qihai Wang, Ivelin Georgiev, Bruce R. Donald, Jingdong Tian:

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Computational and experimental scanning and redesign of nucleic acid proteins. *Sigma Xi, The Scientific Research Society 2009 Student Research Conference.* 2009, The Woodlands in Houston, TX.

5.1. Motivation

Computationally designed and redesigned proteins have been reported. In contrast, comparably fewer numbers of software have been released. We report the release of an open-source cross-platform macromolecular engineering software suite OSPREY: Open Source Protein REdesign for You for the redesign of nucleic acid binding proteins. Functions of the software

suite include protein scanning, redesign, native structure recovery, and molecular rebuilding. The community-wide adoption of OSPREY is anticipated to increase the frequency and diversity of redesigns of nucleic acid binding proteins, ubiquitous and vital to all forms of life.

5.2. Overview

There are a number of significant challenges in computational protein design (197). Access to protein design software should not be one of them. Protein design software is not often readily available or usable. This may be partially due to the state of the software. Development-grade software rarely is distributed outside the expertise and oversight of specific user groups. Another possible reason for the dearth of software may be tied to the conflicts of interests associated with the software. Commercial entities, whether big pharmaceutical corporations or fledgling start-ups often consider software developed in-house as part trade-secret and part competitive advantage.

Software that is readily available may or may not appeal to the end-user. Heuristic algorithms are available, but can be limited in the choice of heuristics and level of control afforded (198,199). Other design routines and libraries are state-of-the-art and powerful, but are not integrated into a relatively user-friendly software suite for end-to-end molecular design (200,201). Some projects have generously established webserver interfaces to automated protein design, which provide some but not complete

customization capabilities, as well as computing clusters, which can be overburdened with potentially many simultaneous end-users (202). While automation in protein design may seem desirable (203), the level and locations of automation can prove otherwise. Still other software are distributed under permissive licenses, but have integrated code from other projects which may not share similar licenses (204). We make the case that many of these qualms can be ameliorated through open-source macromolecular engineering software that is modular and a development community that is correspondingly open and free to exchange ideas.

5.3. Modeling flexibility in nucleic acid binding protein redesign

Macromolecules move. It is desirable for the models in design to move and present different conformations as well, preferably in some principled fashion. Rotamer libraries have continually improved with respect to the high-resolution source crystallographic structures from which they are created (205) and improved modeling using rotamer based-methods (206). Studies of side-chain rearrangements upon ligand binding have shown that, normalized for the number of dihedral bonds, polar amino acids were more flexible than aromatics in interfacial pockets (IPs) (116). Once placed, further flexibility is gained by using energy minimization to relax rotamer side-chain initial geometries (207). This energy minimization may also

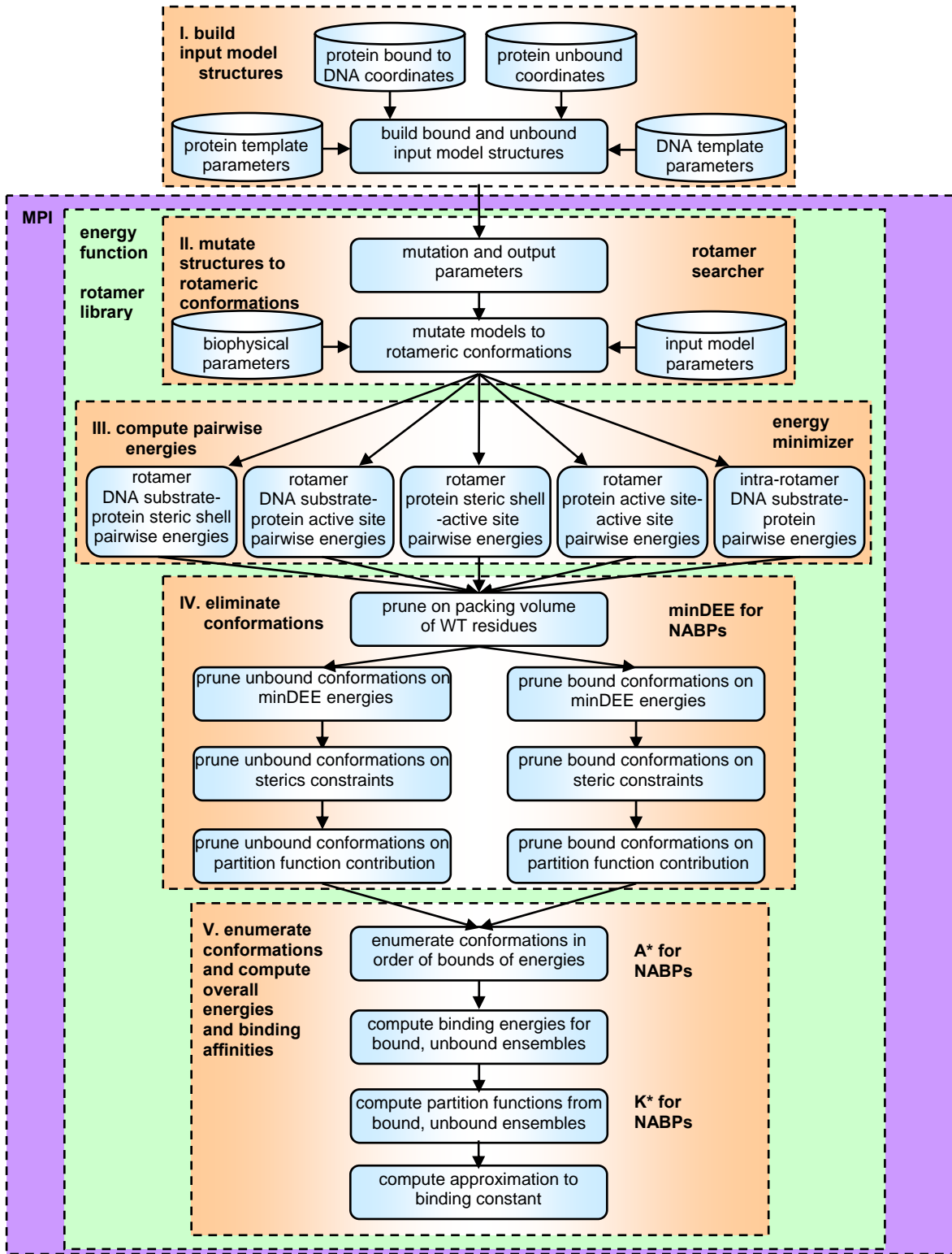
permit more preferable packing and specificity during the design process (208). Even further flexibility can be achieved while maintaining the provable guarantees by moving the protein backbone (209), thus modeling the conformational accommodations that may occur due to mutations to the residue side-chain. While energy functions are continually improving, and energy minimization capabilities reaching to larger and longer molecular space and time scales, optimization methods to estimate and reduce the problem size are being developed (210).

Recognizing the importance of flexibility in design, we have extended the minimized side-chain Dead-End Elimination (minDEE) (112,114), A* gap-free ordered enumeration (115), and K* provably-accurate ensemble-based (124) algorithms to nucleic acid binding proteins in an open-source, cross-platform software suite, OSPREY: Open Source Protein REdesign for You. The software is written in the Java language and uses the standards-compliant message passing interface (MPICH2) and an objected-oriented wrapper (mpiJava) for distributed computation (182-184) (Figure 5-1).

5.4. Computational redesign in practice

OSPREY for nucleic acid binding proteins has been put into practice for four functions: alanine scanning, active site redesign, native structure recovery, and molecular rebuilding of protein and enzyme structures.

Figure 5-1: Open-source software engineering for computational redesign of nucleic acid binding proteins



5.4.1. Scanning

As described in Chapter 4, OSPREY was used to computationally alanine scan R.PvuII residues S81, N140, and N141 for mutation tolerance. Functional assays validated the accuracy of predictions, specifically that residue S81 was tolerant to mutation, while residues N140 and N141 were not tolerant as determined by *in vitro* cleavage assays.

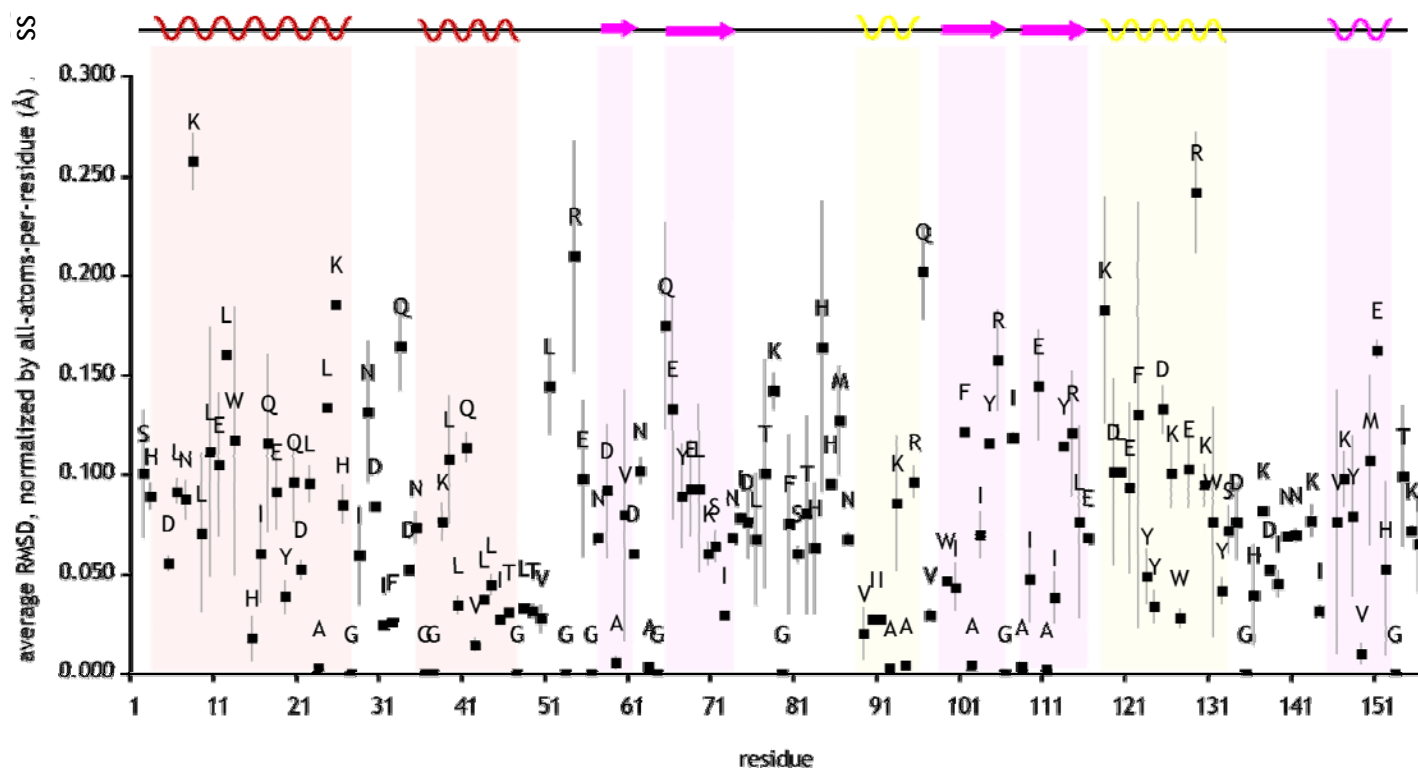
5.4.2. Redesign

As described in Chapter 4, OSPREY was used to redesign R.PvuII residue S81 to all minimized rotamers of all non-proline residues. Time-course functional *in vitro* assays validated the accuracy of predictions. The highest six of the total eight K* ranked predicted mutants all functioned as determined by *in vitro* cleavage assays. Furthermore, the highest K* ranked predicted mutant functioned faster than the wildtype enzyme.

5.4.3. Native structure recovery

A further assessment of computational molecular design algorithms, such as our own, involves their ability to recover what already exists in nature. Termed native recovery, this recapitulation of the existing sequence and structure can be thought of as a redesign towards the natural protein or

Figure 5-2: Native structure recovery of residues in R.PvuII structure

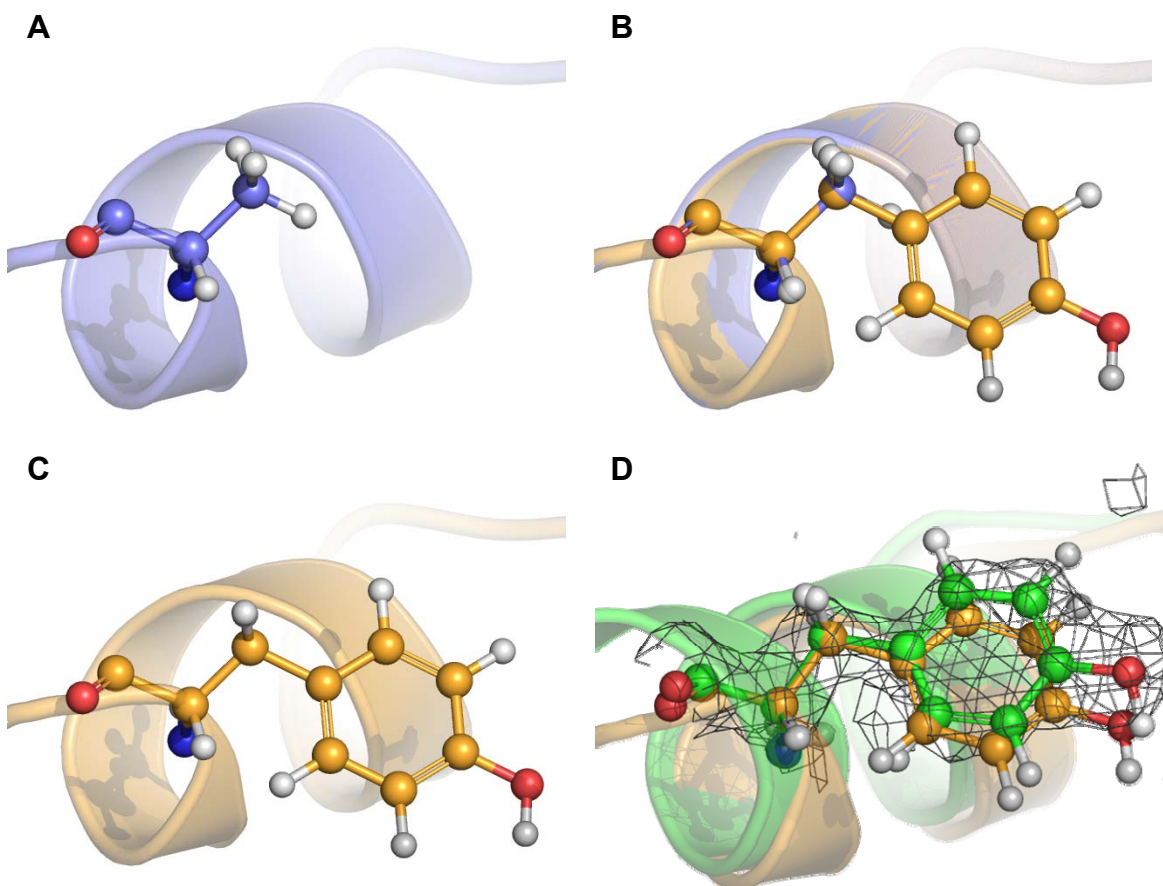


144

Native structure (PDB ID: 1PVI) was recovered for all atoms in all residue side chains as the global minimum energy conformation (GMEC), in terms of small average distance root-mean-square deviation (RMSD). Data point represents mean value of RMSD for the residue found on both monomers of R.PvuII homodimer. Vertical line at each data point represents high value and low value of RMSD for the residue on each monomer of R.PvuII homodimer.

enzyme, given the general framework of the algorithms and the inputs necessary to build the wildtype structure, such as a rotamer library and the sequence of the wildtype protein. OSPREY was used to recover the native structure of R.PvuII at all non-proline residues on both monomers of the homodimer. The bound and unbound 3D structure input models of PvuII were built from source crystallographic structure coordinates for the *holo* form phased at 2.60 Å (PDB accession ID: 1PVI (41)) retrieved from the Protein Data Bank (32) and prepared as described in previous Chapters. The entire protein, cognate and flanking DNA substrate structure was modeled as part of the redesign process. Matching residues on both monomers on the dimer were mutated simultaneously to the wildtype residue, thus predicting rotameric conformations of the wildtype structure. All output model structures were generated and the GMEC for each matching residue identified. Each output model structures were spatially aligned to the input model structure and the distance root-mean-square deviation (RMSD) for each residue was calculated. From this calculation, for each matching residue on the monomers, the average RMSD was computed as well. The average RMSD value was normalized by all-atoms-per-residue to address residue size bias. This average RMSD, normalized by all-atoms-per-residue, was determined to be no greater than 0.3 Å across the entire protein (Figure

Figure 5-3: Molecular rebuilding of R.PvuII placeholder residue A94 to redesigned wildtype residue A94Y



(A) In an earlier crystallographic structure of R.PvuII (PDB ID: 1PVI), the electron density for Y94 was unclear and it was modeled as A94 (purple ball-and-sticks).

(B) Rebuilding the structure from **(A)** to A94Y and the lowest minimum energy side-chain conformation was generated (orange ball-and-sticks).

(C) The rebuilt A94Y lowest minimum energy side-chain conformation (orange ball-and-sticks) from **(B)**.

(D) The rebuilt A94Y lowest minimum energy side-chain conformation (orange ball-and-sticks) aligned to a later crystallographic structure of R.PvuII (PDB ID: 1EYU) having a clearly resolved Y94 residue (green ball-and-sticks) showed spatial and electron density agreement.

5-2). Furthermore, as a computational validation, residues with more dihedral bonds and rotamers in the rotamer library (Appendix A.2) exhibited greater RMSD.

5.4.4. Molecular rebuilding

The ability of our software to natively recover structures accurately suggests the possibility of rebuilding structures that were otherwise left undefined. OSPREY was used to rebuild a placeholder residue, A94, in the aforementioned 3D structure input model of R.PvuII (PDB accession ID: 1PVI (41)). This original authors of the crystallographic structure chose to model residue 94 as an alanine rather than the wildtype tyrosine due to the lack of clarity in the electron density map (Figure 5-3, A). This residue position, originally having alanine as a placeholder, was rebuilt with all flexible rotamers of tyrosine from the rotamer library (Figure 5-3, B). Once again, all output model structures were generated and the lowest energy conformation of A94Y was identified (Figure 5-3, C). As a comparative test, a newer crystallographic structure for which the Y94 residue was adequately resolved to assign a structure to this side-chain (PDB ID: 1EYU) (44) was retrieved from the Protein Data Bank (32). Global space alignment was performed with the CA-C-N_{*i*+1} protein backbone atoms between the output model structure and the structure with Y94. The molecular rebuilt A94Y showed spatial and electron density agreement with the Y92 residue of the newer

crystallographic structure, despite the computations not having prior knowledge of the conformation of the native tyrosine residue (Figure 5-3, D).

5.5. Discussion

The functions demonstrated herein are among some of the many possible using the software suite OSPREY for nucleic acid binding proteins. The open-source, object-oriented framework of this software permits end-users to modify, compile, and run independently. In doing so, further extensions and functions are limited only by the end-user's imagination. Furthermore, the open-source model and reasonably permissive licensing of OSPREY can encourage adoption and improvement through the efforts and good-will of the end-user community.

6. Applications and issues of molecular engineering

*Each individual person is very important.
Each person has tremendous potential.
She or he alone can influence the lives of others
within the communities, nations,
within and beyond her or his own time.*

– Muhammad Yunus
Bangladeshi economist

This chapter has been adapted partially from a manuscript that was joint work with Kuo-Sheng Ma, Ishtiaq Saaem and Jingdong Tian:

Ma K-S., Reza F., Saaem I., Tian J. Versatile surface functionalization of cyclic olefin copolymer (COC) with sputtered SiO₂ thin film for potential BioMEMS applications. *Journal of Materials Chemistry*. 2009, 19: 7914-7920.

partially from a research meeting abstract that was joint work with Jingdong Tian:

Reza F., Tian J. Engineering molecular interactions for targeted therapeutics and technologies. *National Academy of Engineering Grand Challenges National Summit*. 2009, Durham, NC.

and partially from a manuscript that was joint work with Duke University international Genetically Engineered Machines (iGEM) Program 2006:

Reza F., Chandran K., Feltz M., Heinz A., Josephs E., O'Brien P., Van Dyke B., Chung H., Indurkha S., Lakhani N., Lee J., Lin S., Tang N.,

LaBean T., You L., Yuan F., Tian J. Engineering novel synthetic biological systems. *IET Synthetic Biology*. 2007, 1: 48-52.

6.1. Motivation

Engineering biological systems is challenging. Doing so demands intimate understanding of the natural biological system, a multidisciplinary and principled approach, and appreciation of the impact of such actions. The ability to redesign proteins that can act upon nucleic acid substrates is of notable interest, due to their ubiquity and importance in all living organisms.

In basic science and engineering settings, these redesigned proteins can be applied to advancing the state of bionanotechnologies, where DNA enzymes already have a prominent role. Redesigned proteins can also be utilized as components of other biotechnologies, such as the described herein for probing the behaviors of gene-protein circuits.

In medicine, redesigned proteins can be applied as novel therapeutics, such as towards desirable or different signaling and regulation of endogenous genes and proteins or enzyme replacement therapy. As recombinant biologics or biological medicinal products, redesigned proteins can again interface with other therapeutic modalities, such as the strategy described herein of reprogramming of genes for heritable repair or conferral of benefits.

These proposals in biotechnological and therapeutic applications elicit ethical, social, and social issues (ELSI). As molecules that can intimately

modify the code of life in intended as well as deleterious manners, the ethics of nucleic acid binding proteins and enzymes as dual-use technologies are discussed. As useful and novel technologies, these proteins are also within the purview of the legal and industrial sectors, through such avenues as patenting and commercialization. With the emergence and increasing ease of manipulating genes and genomes using nucleic acid binding proteins and other molecules, the social dimensions of do-it-yourself (DIY) and communal hacker biology as well as implications for global health are addressed.

6.2. Bionanotechnologies

Proteins are among biology's most potent nanotechnologies. The diversity of their form and function enables the living world around us. Intrigued and inspired by nature's use of nucleic acid binding proteins, practitioners of the art have adopted these bionanotechnologies to manipulate biology towards various purposes. Researchers have used these proteins as tools of biotechnological discovery and development. By isolating nucleic acid binding proteins from natural systems, observing their behavior, and then creatively connecting this behavior to needs in the laboratory, these proteins have served to copy and thus amplify DNA (211-218), transform DNA into mRNA (219-221), mRNA to DNA (222-226), join DNA together (227-229), repair damaged DNA (230,231), protect or sequester DNA from other molecules (232-235), silence gene expression through RNA interference

(RNAi) (236-238), produce other proteins from mRNA and tRNA (239-243), and much more. Among the functions most desirable and readily applied in the laboratory is the ability to bind and cut short, specific sequences of DNA. This sequence-specific function is undertaken by the workhorses of molecular biology, restriction endonucleases (REases) (16,21,143,144,244-246).

REases have evolved to satisfy the activity and specificity requirements of the host organism. The genes for these enzymes, along with their corresponding methyltransferases (MTases), form restriction-modification (R-M) systems that have been isolated through careful searches of bacterial and archeal genomes. Interestingly, individual R-M systems have highly-specific sequence recognition capacities that are intolerant to degenerate DNA sequences, but collectively these systems are diverse in the sequences they are able to bind (247).

While this diversity exists, the search for natural R-M systems with certain specificities remains unproductive and those that are found are restricted to their evolved levels of enzymatic activity. With the pace of advancements in high-throughput oligonucleotide synthesis and BioMEMS technologies (158,248,249), synthetic gene and even genomes can be artificially produced rather than traditionally cloned from natural sources (250,251). These synthetic genes can be introduced into heterologous cellular expression systems in order to produce proteins. In addition, progress in cell-

free protein expression technologies (190,191,252) permits the production of synthetic proteins from these genes to occur *in vitro*. This is particularly important for the expression of proteins, such as REases, that would be cytotoxic for *in vivo* expression systems that lack sufficient endogenous mechanisms of protection (such as MTases with suitable sequence specificity) and physical barriers (such as nuclear walls) to the activity of these proteins. Analogously, protein-based mechanisms of protection can be created using the aforementioned oligonucleotide synthesis technologies and introduced into the expression system exogenously. Thus, the ability to synthetically produce genes and proteins in conjunction with the ability to mutate genes and produce proteins from nature permits the introduction of greater diversity of specificity and activity than previously available.

The diversity through the redesign of REases for altered specificity and activity was demonstrated in Chapters 3 and 4, respectively. There are a multitude of constructive applications for these redesigned REases. Some of these applications are described as follows.

With altered specificity comes the ability to cleave DNA not previously possible or possible with a less preferable REase. This ability is a boon for laboratory molecular biology, where often experimental conditions would require or prefer such altered specificity. For example, a REase redesigned with a novel specificity would enable new DNA sites to be cleaved and, in

turn, insertion of new genetic material between the cleaved sites to be ligated that was not previously possible. This ability also has consequences for laboratories in various settings. Consider genomic diagnostic tests or in forensic analyses settings, in which DNA profiling through digestion with a REase with redesigned novel specificity can provide further uniquely cut DNA fragments for that single nucleotide polymorphism (SNP). Yet another setting is in genetic counseling and population studies, where a REase with redesigned specificity can digest genomic DNA for restriction fragment length polymorphism analysis. In doing so, closely related sequences that differ at the cognate sequence of the redesigned REase can be distinguished.

With altered activity comes the ability to cleave DNA faster or with fewer non-specific cleavages, or “star” activity, than previously possible. This ability, too, is a boon for molecular biology by reducing incubation time or the amount, and thus cost, of REase needed for complete digestion of the DNA substrate. The faster activity can also enable better monitoring and avoidance of “star” activity, which occurs with prolonged incubations or REases with DNA, and which can obfuscate the reaction results when used in biotechnological and biomedical applications. For similar reasons, mutations to REases that confer reduction in “star” activity while maintaining specificity has been applied (253). Maintaining specificity is vital for utility in sequence-specific biotechnological and biomedical applications, such as

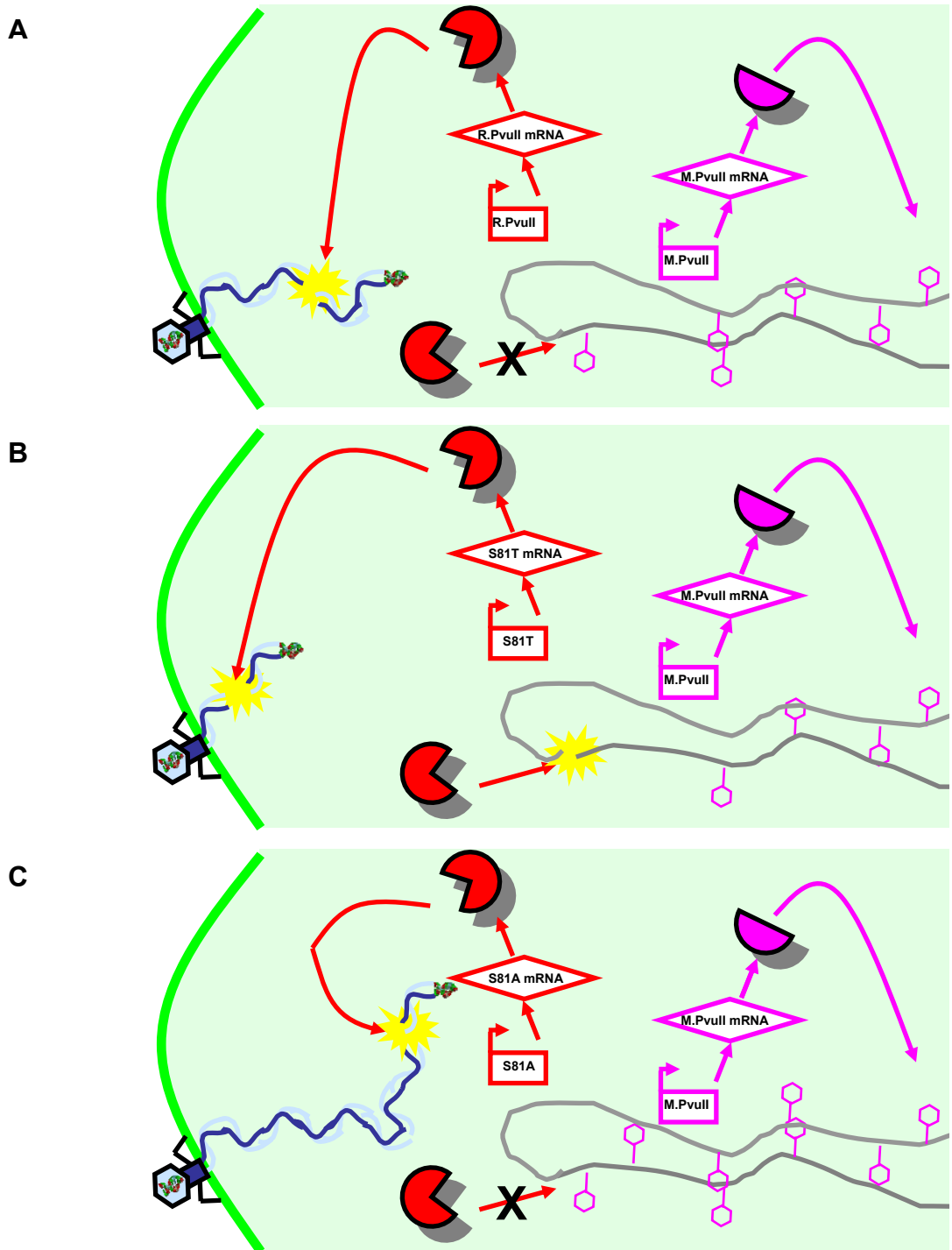
cloning, DNA footprinting, genotyping, and restriction mapping. The combination of both aspects confers the advantages of more rapid, while equally specific, cleavage of the DNA substrate, not found in nature.

The application of redesigned REases also facilitates the ability to test hypotheses about R-M systems themselves. Among the hypotheses on the prevalence of R-M systems are that they are selfish genetic elements that shape the host genome (247). This hypothesis has support from several empirical observations. For example, when a R-M system is introduced into a host, the REase or MTase genes it carries cannot be readily displaced by a plasmid that is incompatible with this introduced system (254) or replaced by DNA that is homologous (255). The incompatible plasmid can also interfere with an introduced or existing R-M system by reducing the concentration of REase and MTase proteins in the cell, leading to “post-segregational host killing.” This killing is hypothesized to occur because while all methylation sites in the host genome must be acted upon by the dwindling number of MTases, only a single un-methylated restriction site is sufficient to be cut by the reduced in number, yet present, REases and cause cell apoptosis (256). Furthermore, the codons used in R-M genes are distinctly different than those used elsewhere in the host genome (257) as well as across several sequenced genomes (258), further implicating that R-M systems are themselves foreign to the host. Another hypothesis on the maintenance of R-

M systems is that they evolved as part of synergistic cellular defense. This hypothesis suggests that prokaryotes evolved to encode REases to defend their genomes from bacterial, viral, and other foreign genetic materials and MTases to protect themselves from their own REases (259,260). This hypothesis is particularly apt to the bacteria and archea in which R-M systems have so far been identified. These hypotheses, and the underlying natural balance that exists between REases and MTases can be tested through REases that maintain sequence specificity but perturb activity from the WT REase. As described in Chapter 5, we have computationally engineered and experimentally evaluated two REases with the requisite activity profiles: R.PvuII-S81T, which has activity faster than R.PvuII-WT, and R.PvuII-S81A, which has activity slower than R.PvuII-WT.

In order to test the selfish genetic elements and synergistic cellular defense hypotheses, cells-based REase expression systems for R.PvuII (186) can be programmed so that synthetic gene-protein circuits interface with the endogenous natural bionanomocular machinery (261). We propose that three synthetic circuits can permit some elucidation of these hypotheses: a circuit consisting of R.PvuII-WT and M.PvuII (Figure 6-1A), a circuit consisting of R.PvuII-S81T and M.PvuII (Figure 6-1B), and a circuit consisting of R.PvuII-S81A and M.PvuII (Figure 6-1C). For testing the post-segregational host

Figure 6-1: Wildtype R.PvuII-WT and redesigned R.PvuII-S81T and R.PvuII-S81A mutants in restriction-modification gene-protein circuits



(A) In a natural restriction-modification gene-protein circuit, the overall rates of the restriction and modification events are tightly regulated so that

the host cell is able to protect its genome through methylation using its own WT M.PvuII before its own WT R.PvuII can cleave the same cognate DNA sequence. The WT R.PvuII is able to cleave the cognate DNA sequence found in the infecting bacteriophage, thus protecting the host.

(B) In a synthetic restriction-modification gene-protein circuit, a mutant R.PvuII-S81T that acts more quickly than the WT R.PvuII may cleave the bacteriophage DNA sequence so that there is less likelihood for phage proliferation, but may cause deregulation in the restriction-modification system so that the same cognate DNA sequence in the host is cleaved before it can be methylated.

(C) In a synthetic restriction-modification gene-protein circuit, a mutant R.PvuII-S81A that acts more slowly than the WT R.PvuII may not cleave the DNA sequence in the host before it can be methylated, but may not also not cleave the bacteriophage DNA sequence as needed so that there is more likelihood for phage proliferation.

killing hypothesis, the faster activity of R.PvuII-S81T may enable it to bind and cleave the host DNA before the M.PvuII has managed to methylate all sites as would be possible in conjunction with the R.PvuII-WT. For testing the cellular defense hypothesis, the slower activity of R.PvuII-S81A may permit greater incorporation of bacteriophage genetic material than would otherwise be possible with R.PvuII-WT. Our preliminary studies indicate that, in fact, cell growth and multiplication is stunted in cultures that are responsible for expressing the R.PvuII mutants when compared to the R.PvuII-WT. Thus, this example highlights how our redesigned REase can further elucidate the workings and timings of natural systems.

6.3. Therapeutics

Proteins are among nature's most sophisticated tools for healing as well as harm. In the case of the former, for example, the first generations of protein-based therapies involved producing and harvesting wildtype proteins in order to apply their natural functions in a therapeutic setting or using monoclonal antibodies to inhibit these natural functions (262). In nanomedicine, antibodies, and non-proteins such as peptides, nucleic acid aptamers, carbohydrates, and small molecules, have been used to differentially target other molecules (263). Protein-based nanotechnology continues to be investigated for therapeutic applications in cancer therapy (264,265), particularly in the first generation modality of antibody therapies

(266-268). Due to the dissemination of new and drug-resistant microorganisms for which anti-microbial agents may be ineffectual or patients' conditions may not permit administration, antibody therapies for infectious diseases are being pursued (269).

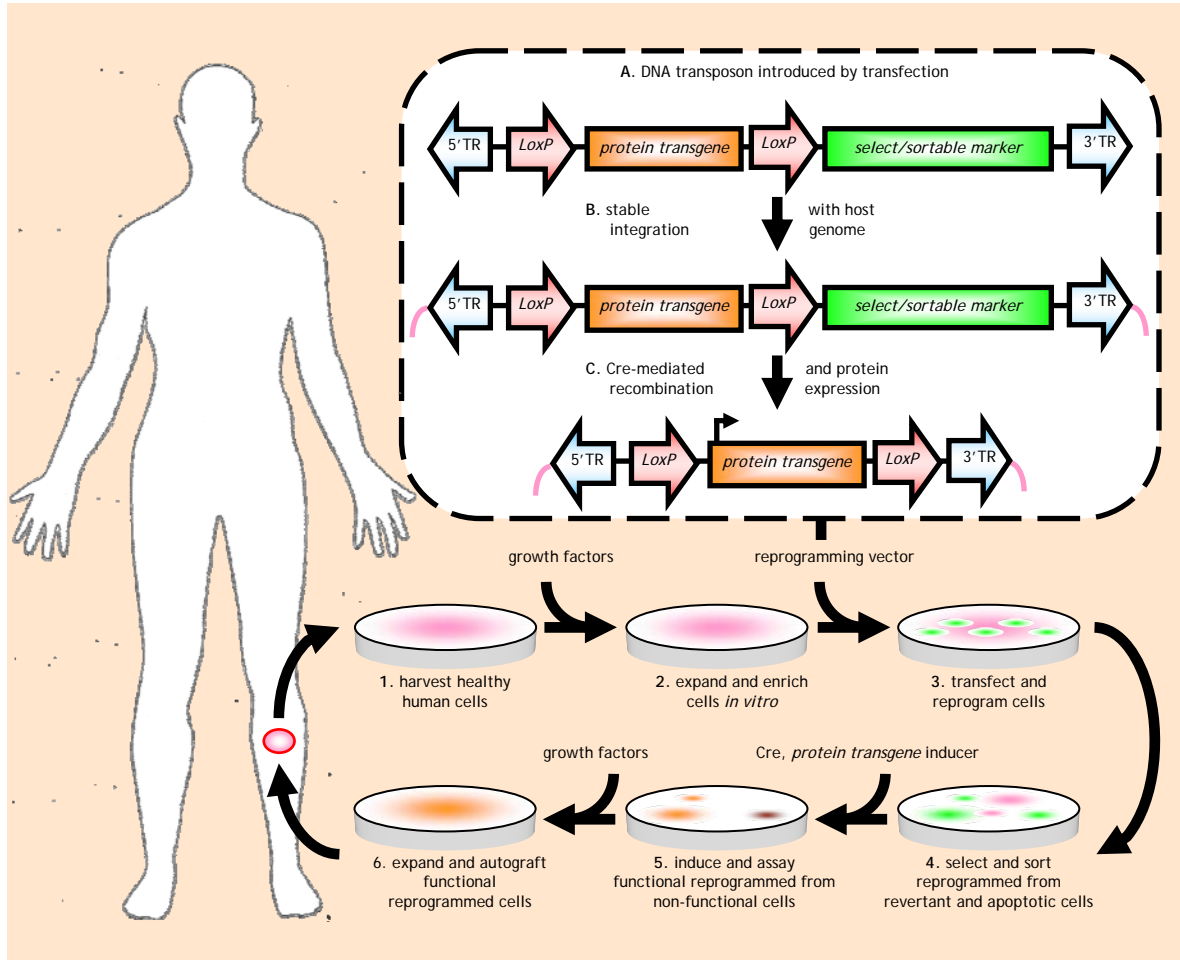
Alternatively, other natural proteins are being targeted for detection and suppression of function rather than their utilization. Telomerases, for example, are essential for the maintenance and immortalization of a subset of cells, including cancer stem cells (270-275). This enzyme is present in more than 85% of cancers but virtually absent from most somatic tissues (276). Thus, the targeting and inhibition of telomerase activity in cancer patients is thought to induce cellular senescence or apoptosis and thus transform the malignancy's immortal phenotype (277).

Natural proteins have also been engineered for therapeutics. This engineering has involved the *in vivo* properties of proteins, such as immunogenicity, affinity, effector functions, and pharmacokinetics (278). Some of these engineering initiatives have taken on a distinct design approach, and marshaled the capabilities of protein modeling and design in particular to modify monoclonal antibodies, cytokines, enzymes and viral fusion inhibitors (279). In intentionally manipulating the physical, chemical and biological properties through structure-based investigation, rational design enables the hypothesis-driven discovery of therapeutic proteins (280).

Among the reported successes in rational design include computational antibody affinity improvement (281), switches in specificity of a non-ribosomal peptide synthetase (181), and design and engineering of an oxygen transport protein (282). Yet, most of these approaches ultimately produce proteins *in vitro* and must be routinely administered due to their transient effects *in vivo* as protein or enzyme replacement therapies.

To produce proteins endogenously in a more permanent manner, the cellular genetics may be reprogrammed by nucleic acid binding proteins and enzymes so that a cell, and its descendants, will express the corrected gene and behavior (Figure 6-2). In this strategy, allografts of cells would reduce the possibility for rejection of the reprogrammed cells after introduction into the host. Among the strategies to introduce such permanent changes is to induce the cellular double strand break (DSB) and repair by non-homologous end joining (NHEJ) machinery (283,284) while providing a DNA template containing the genetic changes to be introduced into the host genome. This strategy has been achieved using triplex forming oligonucleotides (TFOs) (285,286) to elicit the machinery to correct the beta-globin gene in a heritable fashion (176,287,288). A transgene encoding a redesigned REase with more potent or altered specificity would provide more efficient DSB induction than is currently possible using TFOs. This greater number of DSBs may increase the likelihood of proper NHEJ with the provided template DNA containing

Figure 6-2: Nucleic acid binding proteins and enzymes as gene reprogramming therapeutics



Autograft of cells reprogrammed with REase transgene confers host with targeted protein therapy of deleterious human genes or against invading viral genomes. Upon introduction as a genome therapeutic via autograft cells they may induce the endogenous double-strand break (DSB) and repair machinery of human genes or protection against viral genes.

the genetic changes to be incorporated. In addition, taking a cue from nature, eukaryotic cells could encode and use REases without the need for the corresponding MTases (since they have nuclear walls to protect their genomes from cleavage). These REases would act as cellular sentinels against viral genes in the cytoplasm, much the same way bacteria and archaea use them in synergistic cellular defense. Redesigned REases can thus be applied against some of the prevalent viral scourges, including HIV (289). In general, reprogramming cells using redesigned nucleic acid binding proteins and enzymes provides promising prospects for complementing or supplementing the native physiology in a heritable fashion.

6.4. Ethical, legal, and social issues (ELSI)

The ease and access with which genes and genomes can be created and manipulated with nucleic acid binding proteins and enzymes raises ethical, legal, and social issues (ELSI) that go beyond the laboratory. DNA synthesis and sequencing productivity has been increasing at an insatiable pace (248). The proliferation and progress in biotechnology has enabled the average estimated time required to obtain protein structures, including isolation/production, crystallization, data collection, and model building, to reduce from a decade to less than a year in person-years (290).

The ethical issues with redesigning nucleic acid binding proteins and enzymes involve concerns for the unpredictable consequences that altering

these proteins, and thus the balance they maintain, may have in all living beings. In this regard, these proteins can be considered dual-use technologies, with the potential for benefit, such as the aforementioned bionanotechnological and therapeutic examples, or harm. Adding to the social fabric, the relatively automated and cost-effective means of producing genes and proteins *de novo* has fostered a do-it-yourself (DIY) and hacker biology cultures to create novel synthetic biological systems (291-293).

Legal quandaries regarding the products of these cultures include intellectual property and ownership rights (294-296) and the interaction of these cultures with the proprietary research and development interests (297), such as in the case of drug discovery (298). These communities need not be shunned or regulated but rather supported and encouraged to work collaboratively, openly, and constructively.

The social aspects can extend more broadly, with implications for global health and well-being. Given the aforementioned uses of nucleic acid binding proteins as gene-protein circuits, intimately tied to and affecting the host cell physiology, and REase reprogrammed cells, which need not require constant reintroduction, it is not far from the imagination for such bionanotechnology and heritable therapeutics to be deployed in corners of the world where constant health monitoring and administration to humans and livestock are impracticalities or luxuries.

6.5. Discussion

The redesign of nucleic acid binding proteins and enzymes and their applications in bionanotechnology and therapeutics are ongoing, forthcoming, and realizable. No longer are the province of speculation (299-301) or science fiction (302-308), nature's nanotechnology and therapeutics in the form of proteins are being molecular engineered as-is, whole or part, or redesigned as we have done, in order to perform novel and different tasks. An open approach to dealing with these engineered molecules, just as in the open-source practice of sharing and improving the tools that design them, as expounded upon in Chapter 5, will enable more effective and safer adoption for all.

7. Conclusion

*For I dipped into the future,
far as human eye could see,
saw the vision of the world,
and all the wonder that would be.*

– Alfred Tennyson
British poet

7.1. Contributions of this dissertation

In this dissertation the computational molecular engineering of nucleic acid binding protein and enzymes was presented. The contributions of this dissertation commence with modeling interactions among DNA and the proteins that bind and catalyze them, proceeds to engineering these interactions at single- and multiple-conformation means, disseminates the computational tools to perform similar designs by a community of computational molecular engineers as well as the uninitiated, and presents the applications and issues that these molecular tools and technologies can have on society.

In Chapter 2, principles of natural systems involving DNA, DNA binding proteins, and restriction endonucleases (REases), and their interactions were presented. Through the analysis of model of natural systems a number of interesting conclusions were drawn. Gene-protein circuits were modeled using both analytical as well as stochastic methods. Both methods identified a critical frequency, f_c , of input signal, f_{in} ,

transmission from the beginning of the circuit beyond which the frequency signal reported, f_{out} , at the end of the circuit became corrupted. Furthermore, DNA-protein structures were modeled using structural bioinformatics and 3D visualization tools. This investigation quantified the degree of distortion of DNA substrates that were bound to restriction endonucleases (REases). The crystallographic bound DNA substrates modeled with the tools demonstrated varying degrees of distortion and deviation from canonical B-DNA. The structural aspects of DNA-protein structures were pursued in many of the latter Chapters.

In Chapter 3, the single-conformation engineering of nucleic acid binding proteins from models of DNA-protein structures was investigated. Computational filtering and biological focusing approaches were examined, and a coupled computational and biological approach was further implemented. Focusing and filtering at different structural levels of the chosen REase molecular system, R.PvuII, predicted a mutant with altered substrate sequence specificity which was validated through the development of cell survival and enzymatic activity assays.

In Chapter 4, the multiple-conformation engineering of nucleic acid binding proteins from models of DNA-protein structures was studied. Modeling was done in depth of the REase R.PvuII and a region critical to binding and cleavage was chosen for redesign. The state-of-the-art provably-

accurate design algorithms, minimized side-chain Dead-End Elimination (minDEE), A* and K* were extended to proteins and enzymes that bind nucleic acids. In doing so, DNA was among the largest known protein substrates simulated using these algorithms. Computational alanine scanning in the region predicted residues that were tolerant or intolerant to mutation, of which all were experimentally shown to be as predicted. Redesign of the tolerant residue predicted eight mutant proteins with preference for the cognate DNA substrate, of which the top ranked six bound and cleaved as predicted. The top ranked redesigned mutant R.PvuII-S81T out-performs the natural wildtype protein, by consistently cleaving DNA substrates faster under the same buffer, cofactor, and DNA substrate conditions while maintaining substrate sequence specificity. Given that often a single variation in the DNA sequence encoding a restriction endonuclease or the environment in which it functions can result in over millions-fold reductions in activity or degeneration in specificity (86,309-312), and given that native sequences are already nearly optimal (313), it is quite notable that our top-ranked prediction of R.PvuII, among the most challenging to redesign given that it is the smallest REase known, was still successful in both increasing activity as well as maintaining specificity.

In Chapter 5, open-source molecular engineering was advocated. An overview of our software release for the scanning and redesign of nucleic acid

binding proteins was presented. Modeling flexibility in nucleic acid binding protein redesign by our software was among the novel aspects described. Additional applications of the software, such as native structure recovery and molecular rebuilding, were demonstrated. It is hoped that this open-source software will enable the molecular engineering community to proliferate and the number and frequency of molecular design efforts and successes to increase.

In Chapter 6, the applications and issues of molecular engineering were considered. Applications of designed nucleic acid binding proteins and enzymes in bionanotechnologies and in therapeutics were presented. In bionanotechnologies, redesigned restriction endonucleases, such as the ones we created as part of this dissertation, were proposed to facilitate further discovery and biotechnological development. In addition, they were also proposed for investigating hypotheses about the origins and behavior of restriction endonucleases themselves. A series of restriction-modification gene-protein circuits were proposed to examine the prevailing hypotheses of restriction modification (R-M) systems as selfish genetic elements as well as for synergistic cellular defense. As therapeutics, redesigned nucleic acid binding proteins can be applied in a number of treatment modalities, including therapeutics against cancer, aging, and viral infections. Redesigned nucleic acid enzymes were proposed to bind deleterious genes and

induce the endogenous DNA repair machinery and, in conjunction with the introduced template of the corrected gene, reprogram the host's own cells. Borrowing from the behavior found in nature, redesigned REases were also proposed as effective cellular defense mechanisms for cells with nuclear walls, such as eukaryotic cells, without the necessity for matching MTases. The ethical, legal, and social issues (ELSI) of redesigning nucleic acid binding proteins were discussed. In the context of dual-use technologies, conclusions drawn included the observation of the unfettered pace of engineered genomic and proteomic technologies and the recommendation that open access and utilization of information can facilitate safe applications as well as provide security. The proliferation of means for engineering biology has fostered the emergence of a DIY and hacker culture in biology in individual and small group settings that deserves note. On a global health scale, the potential for engineered biology to modulate and maintain well-being with minimal human intervention is recognized.

Through the works in this dissertation, it is hoped that harnessing the power of computation and models can further our capacity to extend the frontier of atomic biology through molecular design. The contributions made herein to computationally design novel proteins as intended with the advancements in materializing these designs through *de novo* molecular synthesis and expression technologies permits us to glean into the workings

of living molecular machines that were at times only within the purview of nature. In doing so, we promote and benefit from the intelligent evolution of molecular design.

7.2. Intelligent evolution of molecular design

Like the evolved molecules that they are entrusted to help engineer, molecular design methods are evolving at a staggering pace due to increasing interest, necessity, and capabilities.

Progress in computational protein and enzyme design has been occurring on many fronts, including improved energy functions and better search and optimization procedures (314). This progress has been fueled by increases in the number of molecular structures determined and their deposition to and availability from public databases (31). Scientific computing tools have matured further and provided processing power at an exponential rate (315) to enable more sophisticated methods to be implemented and designs to be attempted. The processing capacity has also been increase by distributed processing tasks across computers, as is the case of some of the work described in this dissertation and in the work of others (316-323).

Synthetic experimental technologies, too, have emerged in order to facilitate the creation of DNA and proteins not already in existence and they

are being used in projects that affect the molecular status quo of society (324-331) and the individual (332-334).

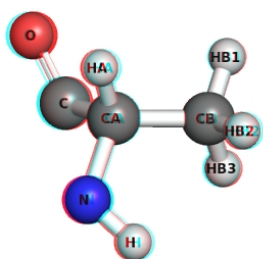
With advancements on so many fronts, the conversations, conversations, and cooperation continue on the scientific as well as societal repercussions of molecular engineering, much like those from the last century on recombinant DNA technology (335,336) and emerging DNA analytical methods (337-339).

Thus, computational molecular engineering is coming of age and continues to be a driving force for scientific, technological, and social progress...

Appendix

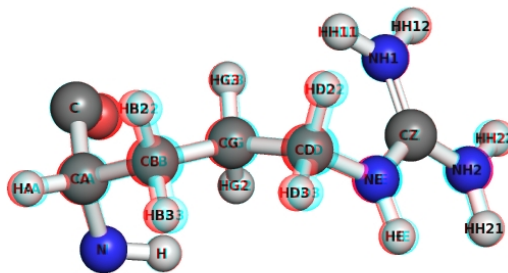
A.1. Amino acid templates

alanine template



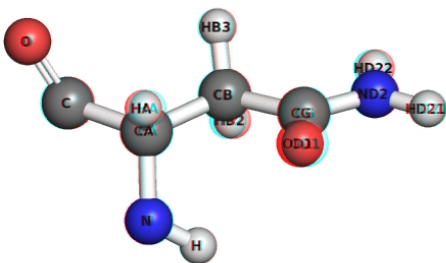
atom names for
input model residue name: **ALA**
atom count: **10 atoms**

arginine template



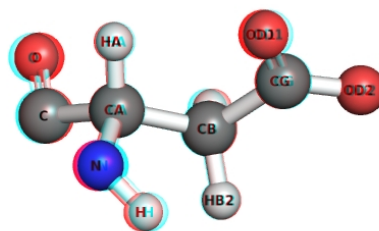
atom names for
input model residue name: **ARG**
atom count: **24 atoms**

asparagine template



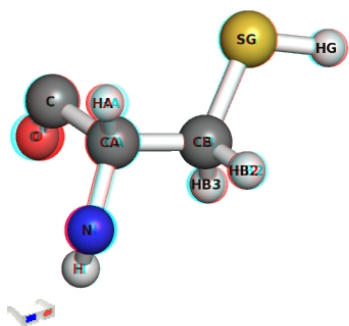
atom names for
input model residue name: **ASN**
atom count: **14 atoms**

aspartic acid template



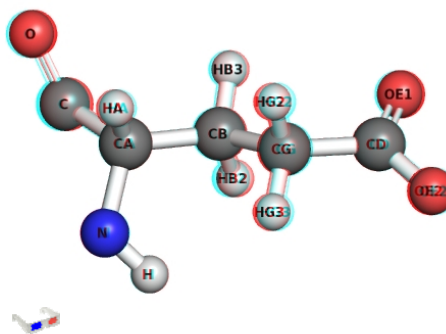
atom names for
input model residue name: **ASP**
atom count: **12 atoms**

cysteine template



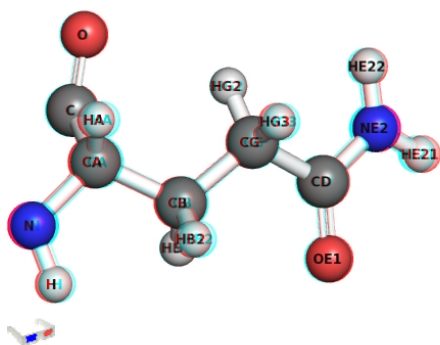
atom names for
input model residue name: **CYS**
atom count: **11 atoms**

glutamic acid template



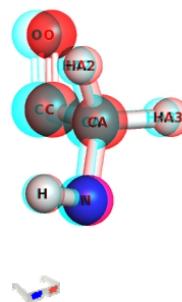
atom names for
input model residue name: **GLU**
atom count: **15 atoms**

glutamine template



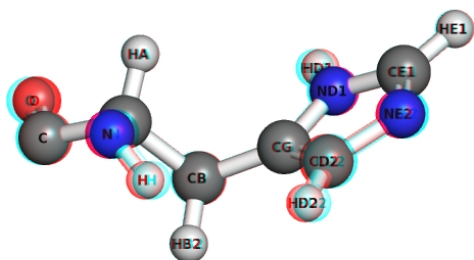
atom names for
input model residue name: **GLN**
atom count: **17 atoms**

glycine template

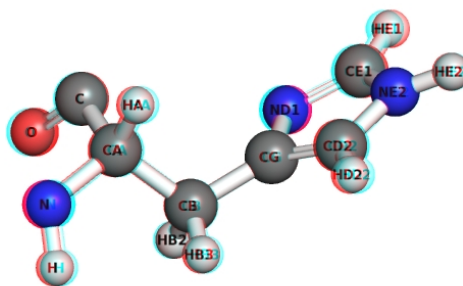


atom names for
input model residue name: **GLY**
atom count: **7 atoms**

histidine templates



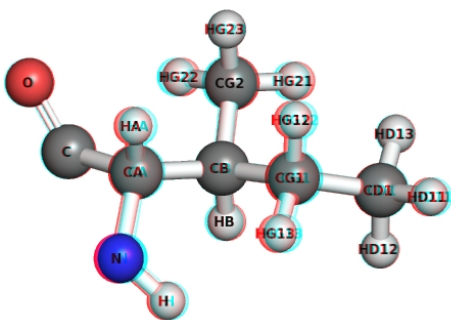
atom names for
input model residue name: **HID**
atom count: **17 atoms**



atom names for
input model residue name: **HIE**
atom count: **17 atoms**

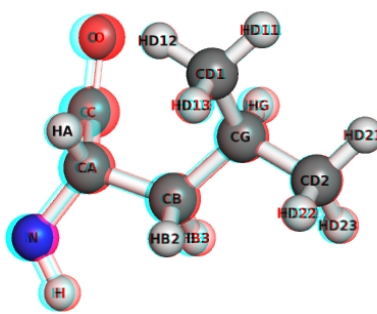
Note: if hydrogen atoms are present on both imidazole nitrogen atoms, i.e. **HD1** on **ND1** and **HE2** on **NE2**, then input model residue name: **HIP**)

isoleucine template



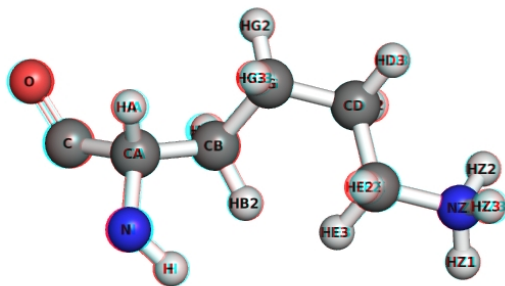
atom names for
input model residue name: **ILE**
atom count: **19 atoms**

leucine template



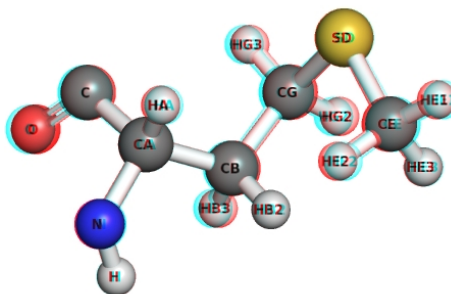
atom names for
input model residue name: **LEU**
atom count: **19 atoms**

lysine template



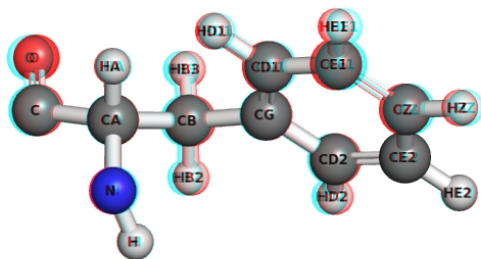
atom names for
input model residue name: **LYS**
atom count: **22 atoms**

methionine template



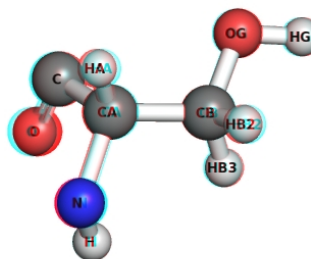
atom names for
input model residue name: **MET**
atom count: **17 atoms**

phenylalanine template



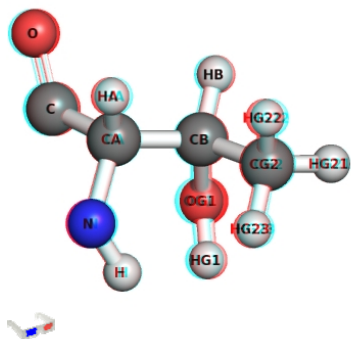
atom names for
input model residue name: **PHE**
atom count: **20 atoms**

serine template



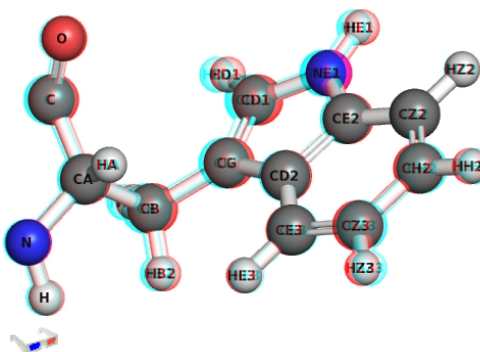
atom names for
input model residue name: **SER**
atom count: **11 atoms**

threonine template



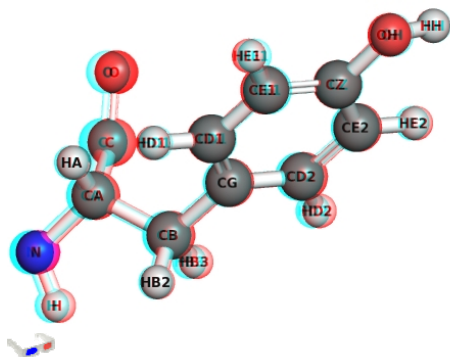
atom names for
input model residue name: **THR**
atom count: **14 atoms**

tryptophan template



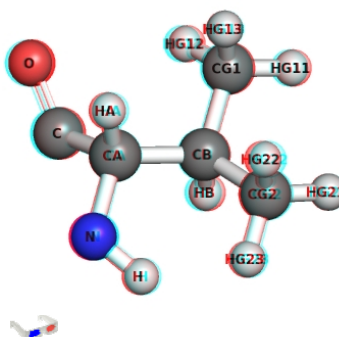
atom names for
input model residue name: **TRP**
atom count: **24 atoms**

tyrosine template



atom names for
input model residue name: **TYR**
atom count: **21 atoms**

valine template



atom names for
input model residue name: **VAL**
atom count: **16 atoms**

Modeled from source crystallographic structures for REase R.PvuII (PDB IDs: 1PVI, 1H56) and phenylalanine activating domain of gramicidin synthetase (PDB ID: 1AMU).

A.2. Amino acid rotamer library

amino acid	number of rotamers	number of dihedrals	atom name of dihedral angles	X ₁ angle (°)	X ₂ angle (°)	X ₃ angle (°)	X ₄ angle (°)
alanine	0 (1 struc.)	0					
arginine	34	4	X ₁ : N-CA-CB-CG	62	180	65	85
			X ₂ : CA-CB-CG-CD	62	180	65	-175
			X ₃ : CB-CG-CD-NE	62	180	180	85
			X ₄ : CG-CD-NE-CZ	62	180	180	180
				62	180	180	-85
				62	180	-65	175
				62	180	-65	-85
				-177	65	65	85
				-177	65	65	-175
				-177	65	180	85
				-177	65	180	180
				-177	180	65	85
				-177	180	65	-175
				-177	180	65	-105
				-177	180	180	85
				-177	180	180	180
				-177	180	180	-85
				-177	180	-65	105
				-177	180	-65	175
				-177	180	-65	-85
	-67	180	65	85			
	-67	180	65	-175			
	-67	180	65	-105			
	-67	180	180	85			
	-67	180	180	180			
	-67	180	180	85			
	-67	180	-65	105			
	-67	180	-65	175			
	-67	-167	-65	-85			
	-62	-68	180	85			
	-62	-68	180	180			
	-62	-68	180	-85			
	-62	-68	-65	175			
	-62	-68	-65	-85			
asparagine	7	2	X ₁ : N-CA-CB-CG	62	-10		
			X ₂ : CA-CB-CG-OD1	62	30		
				-174	-20		
				-177	30		
				-65	-20		
				-65	-75		
	-65	120					
aspartic acid	5	2	X ₁ : N-CA-CB-CG	62	-10		
			X ₂ : CA-CB-CG-OD1	62	30		
				-177	0		
				-177	65		
				-70	-15		

amino acid	number of rotamers	number of dihedrals	atom name of dihedral angles	X1 angle (°)	X2 angle (°)	X3 angle (°)	X4 angle (°)
cysteine	3	1	X1: N-CA-CB-SG	62 -177 -65			
glutamic acid	8	3	X1: N-CA-CB-CG X2: CA-CB-CG-CD X3: CB-CG-CD-OE1	62 70 -177 -177 -177 -65 -67 -65	180 -80 65 180 -80 85 180 -65	-20 0 10 0 -25 0 -10 -40	
glutamine	9	3	X1: N-CA-CB-CG X2: CA-CB-CG-CD X3: CB-CG-CD-OE1	62 70 -177 -177 -177 -65 -67 -65 -65	180 -75 65 65 180 85 180 -65 -65	20 0 -100 60 0 0 -25 -40 100	
glycine	0 (1 struc.)	0					
histidine	8	2	X1: N-CA-CB-CG X2: CA-CB-CG-ND1	62 62 -177 -177 -177 -65 -65 -65	-75 80 -165 -80 60 -70 165 80		
isoleucine	7	2	X1: N-CA-CB-CG1 X2: CA-CB-CG1-CD1	62 62 -177 -177 -65 -65 -57	100 170 66 165 100 170 -60		
leucine	5	2	X1: N CA CB CG X2: CA CB CG CD1	62 -177 -172 -85 -65	80 65 145 65 175		

amino acid	number of rotamers	number of dihedrals	atom name of dihedral angles	X1 angle (°)	X2 angle (°)	X3 angle (°)	X4 angle (°)
lysine	27	4	X1: N-CA-CB-CG	62	180	68	180
			X2: CA-CB-CG-CD	62	180	180	65
			X3: CB-CG-CD-CE	62	180	180	180
			X4: CG-CD-CE-NZ	62	180	180	-65
				62	180	-68	180
				-177	68	180	65
				-177	68	180	180
				-177	68	180	-65
				-177	180	68	65
				-177	180	68	180
				-177	180	180	65
				-177	180	180	180
				-177	180	180	-65
				-177	180	-68	180
				-177	180	-68	-65
				-90	68	180	180
				-67	180	68	65
				-67	180	68	180
				-67	180	180	65
			methionine	13	3	X1: N-CA-CB-CG	62
X2: CA-CB-CG-SD	62	180				-75	
X3: CB-CG-SD-CE	-177	65				75	
	-177	65				180	
	-177	180				75	
	-177	180				180	
	-177	180				-75	
	-67	180				75	
	-67	180				180	
	-67	180				-75	
	-65	-65				103	
	-65	-65				180	
	-65	-65				-70	
phenyl alanine	4	2	X1: N CA CB CG	62	90		
			X2: CA CB CG CD1	-177	80		
				-65	-85		
				-65	-30		
serine	3	1	X1: N-CA-CB-OG	62			
				-177			
				-65			

amino acid	number of rotamers	number of dihedrals	atom name of dihedral angles	X1 angle (°)	X2 angle (°)	X3 angle (°)	X4 angle (°)
threonine	3	1	X1: N-CA-CB-OG1	62 -175 -65			
tryptophan	7	2	X1: N-CA-CB-CG X2: CA-CB-CG-CD1	62 62 -177 -177 -65 -65 -65	-90 90 -105 90 -90 -5 95		
tyrosine	4	2	X1: N-CA-CB-CG X2: CA-CB-CG-CD1	62 -177 -65 -65	90 80 -85 -30		
valine	3	1	X1: N-CA-CB-CG1	63 175 -60			

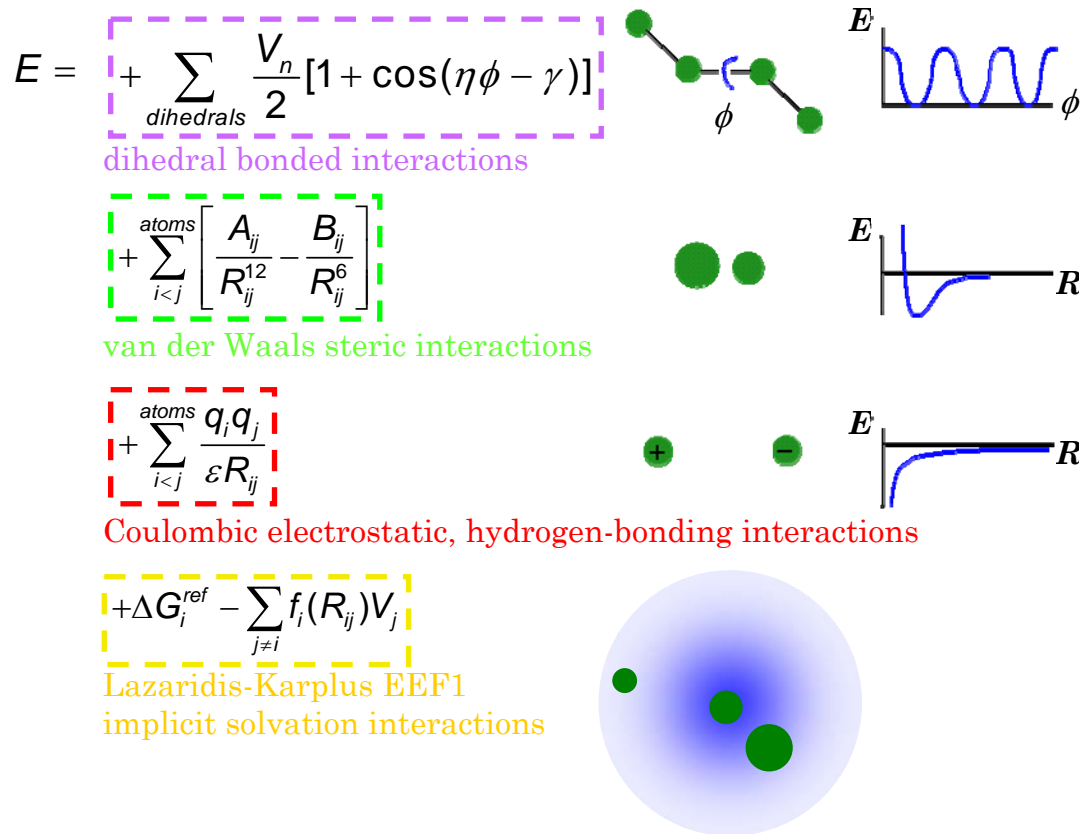
adapted from (103).

A.3. Amino acid rotamer volumes

amino acid	volume (Å ³)					
alanine	61.84375					
arginine	131.53125	131.17188	131.42188	131.09375	131.29688	131.4375
	131.14062	131.26562	131.375	131.10938	131.4375	130.96875
	131.26562	131.32812	131.46875	131.1875	131.57812	131.6875
	131.375	131.28125	131.29688	131.625	131.45312	131.26562
	131.51562	131.17188	131.20312	131.4375	131.1875	131.71875
	130.82812	130.98438	131.5	131.625		
asparagine	88.171875	88.296875	88.15625	88.953125	88.78125	88.890625
	88.765625					
aspartic acid	83.125	83.453125	83.46875	83.84375	84.109375	
cysteine	80.234375	80.171875	80.546875			
glutamic acid	96.96875	96.1875	96.796875	97.296875	96.765625	97.015625
	97.140625	97.546875				
glutamine	102.46875	101.109375	102.265625	102.296875	102.4375	102.3125
	102.34375	102.59375	102.109375			
glycine	48.109375					
histidine	110.453125	110.25	110.71875	110.703125	110.96875	110.6875
	110.859375	110.640625				
isoleucine	102.59375	103.09375	102.859375	102.984375	102.90625	102.734375
	102.671875					
leucine	101.90625	103.0625	102.984375	102.734375	103.125	
lysine	114.40625	114.359375	114.078125	114.4375	114.359375	114.25
	114.40625	114.328125	114.453125	114.40625	114.40625	114.59375
	114.421875	114.453125	114.609375	114.59375	114.375	114.640625
	114.40625	114.546875	114.515625	114.46875	114.65625	114.546875
	114.640625	114.546875	114.640625			
methionine	108.0	107.828125	107.78125	107.671875	108.265625	107.890625
	108.15625	108.234375	108.171875	108.015625	107.5	107.96875
	107.640625					
phenyl alanine	126.015625	125.796875	126.03125	125.90625		
serine	68.890625	68.921875	68.65625			
threonine	82.6875	82.34375	82.28125			
tryptophan	151.32812	150.98438	151.10938	151.26562	151.67188	151.5625
	151.65625					
tyrosine	132.23438	132.40625	132.35938	132.23438		
valine	89.203125	89.1875	89.234375			

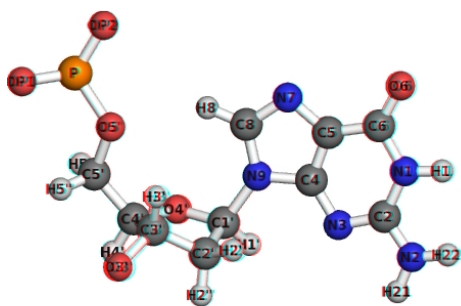
computed from Appendix A.2. Amino acid rotamer library

A.4. Energy function



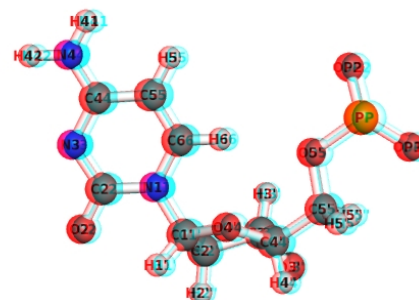
A.5. Nucleic acid templates

guanine template



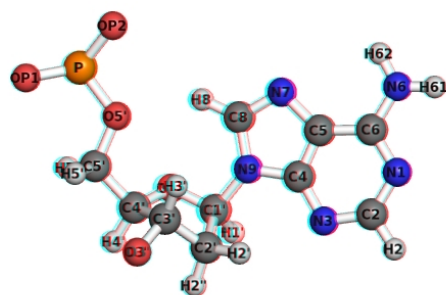
atom names for
input model residue name: **DG**
atom count: **33 atoms**

cytosine template



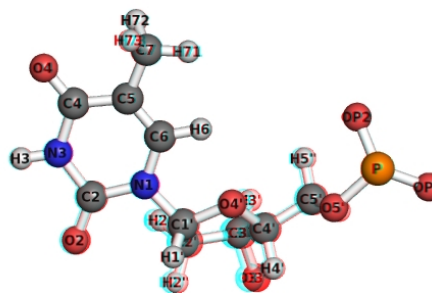
atom names for
input model residue name: **DC**
atom count: **30 atoms**

adenine template



atom names for
input model residue name: **DA**
atom count: **32 atoms**

thymine template



atom names for
input model residue name: **DT**
atom count: **32 atoms**

Modeled from source crystallographic structure for R.PvuII (PDB IDs: 1PVI).

References

- (1) Gilbert W. Origin of Life - the RNA World. *Nature*. 1986, 319: 618-618.
- (2) Avery O. T., MacLeod C. M., McCarty M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from *Pneumococcus* Type III. *Journal of Experimental Medicine*. 1944, 79: 137-158.
- (3) Watson J. D., Crick F. H. C. Molecular Structure of Nucleic Acids - a Structure for Deoxyribose Nucleic Acid. *Nature*. 1953, 171: 737-738.
- (4) Chandrasekaran R., Arnott S. The structure of B-DNA in oriented fibers. *Journal of Biomolecular Structure and Dynamics*. 1996, 13: 1015-1027.
- (5) Schleif R. DNA-Binding by Proteins. *Science*. 1988, 241: 1182-1187.
- (6) Laskowski R. A., Luscombe N. M., Swindells M. B., Thornton J. M. Protein clefts in molecular recognition and function. *Protein Science*. 1996, 5: 2438-2452.
- (7) Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J. P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J. C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R. H., Wilson R. K., Hillier L. W., McPherson J. D., Marra M. A., Mardis E. R., Fulton L. A., Chinwalla A. T., Pepin K. H., Gish W. R., Chissoe S. L., Wendl M. C., Delehaunty K. D., Miner T. L., Delehaunty A.,

Kramer J. B., Cook L. L., Fulton R. S., Johnson D. L., Minx P. J., Clifton S. W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J. F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs R. A., Muzny D. M., Scherer S. E., Bouck J. B., Sodergren E. J., Worley K. C., Rives C. M., Gorrell J. H., Metzker M. L., Naylor S. L., Kucherlapati R. S., Nelson D. L., Weinstock G. M., Sakaki Y., Fujiyama A., Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T., Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E., Robert C., Wincker P., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A., Yang H. M., Yu J., Wang J., Huang G. Y., Gu J., Hood L., Rowen L., Madan A., Qin S. Z., Davis R. W., Federspiel N. A., Abola A. P., Proctor M. J., Myers R. M., Schmutz J., Dickson M., Grimwood J., Cox D. R., Olson M. V., Kaul R., Raymond C., Shimizu N., Kawasaki K., Minoshima S., Evans G. A., Athanasiou M., Schultz R., Roe B. A., Chen F., Pan H. Q., Ramser J., Lehrach H., Reinhardt R., McCombie W. R., de la Bastide M., Dedhia N., Blocker H., Hornischer K., Nordsiek G., Agarwala R., Aravind L., Bailey J. A., Bateman A., Batzoglu S., Birney E., Bork P., Brown D. G., Burge C. B., Cerutti L., Chen H. C., Church D., Clamp M., Copley R. R., Doerks T., Eddy S. R., Eichler E. E., Furey T. S., Galagan J., Gilbert J. G. R., Harmon C., Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W. H., Johnson L. S., Jones T. A., Kasif S., Kasprzyk A., Kennedy S., Kent W. J., Kitts P., Koonin E. V., Korf I., Kulp D., Lancet D., Lowe T. M., McLysaght A., Mikkelsen T., Moran J. V., Mulder N., Pollara V. J., Ponting C. P., Schuler G., Schultz J. R., Slater G., Smit A. F. A., Stupka E., Szustakowki J., Thierry-Mieg D., Thierry-Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf Y. I., Wolfe K. H., Yang S. P., Yeh R. F., Collins F., Guyer M. S., Peterson J., Felsenfeld A., Wetterstrand K. A., Patrinos A., Morgan M. J., Conso I. H. G. S. Initial sequencing and analysis of the human genome. *Nature*. 2001, 409: 860-921.

- (8) Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., Gocayne J. D., Amanatides P., Ballew R. M., Huson D. H., Wortman J. R., Zhang Q., Kodira C. D., Zheng X. Q. H., Chen L., Skupski M., Subramanian G., Thomas P. D., Zhang J. H., Miklos G. L. G., Nelson C., Broder S., Clark A. G., Nadeau C., McKusick V. A., Zinder N., Levine A. J., Roberts R. J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S.,

Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z. M., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A. E., Gan W., Ge W. M., Gong F. C., Gu Z. P., Guan P., Heiman T. J., Higgins M. E., Ji R. R., Ke Z. X., Ketchum K. A., Lai Z. W., Lei Y. D., Li Z. Y., Li J. Y., Liang Y., Lin X. Y., Lu F., Merkulov G. V., Milshina N., Moore H. M., Naik A. K., Narayan V. A., Neelam B., Nusskern D., Rusch D. B., Salzberg S., Shao W., Shue B. X., Sun J. T., Wang Z. Y., Wang A. H., Wang X., Wang J., Wei M. H., Wides R., Xiao C. L., Yan C. H., Yao A., Ye J., Zhan M., Zhang W. Q., Zhang H. Y., Zhao Q., Zheng L. S., Zhong F., Zhong W. Y., Zhu S. P. C., Zhao S. Y., Gilbert D., Baumhueter S., Spier G., Carter C., Cravchik A., Woodage T., Ali F., An H. J., Awe A., Baldwin D., Baden H., Barnstead M., Barrow I., Beeson K., Busam D., Carver A., Center A., Cheng M. L., Curry L., Danaher S., Davenport L., Desilets R., Dietz S., Dodson K., Doup L., Ferriera S., Garg N., Gluecksmann A., Hart B., Haynes J., Haynes C., Heiner C., Hladun S., Hostin D., Houck J., Howland T., Ibegwam C., Johnson J., Kalush F., Kline L., Koduru S., Love A., Mann F., May D., McCawley S., McIntosh T., McMullen I., Moy M., Moy L., Murphy B., Nelson K., Pfannkoch C., Pratts E., Puri V., Qureshi H., Reardon M., Rodriguez R., Rogers Y. H., Romblad D., Ruhfel B., Scott R., Sitter C., Smallwood M., Stewart E., Strong R., Suh E., Thomas R., Tint N. N., Tse S., Vech C., Wang G., Wetter J., Williams S., Williams M., Windsor S., Winn-Deen E., Wolfe K., Zaveri J., Zaveri K., Abril J. F., Guigo R., Campbell M. J., Sjolander K. V., Karlak B., Kejariwal A., Mi H. Y., Lazareva B., Hatton T., Narechania A., Diemer K., Muruganujan A., Guo N., Sato S., Bafna V., Istrail S., Lippert R., Schwartz R., Walenz B., Yooseph S., Allen D., Basu A., Baxendale J., Blick L., Caminha M., Carnes-Stine J., Caulk P., Chiang Y. H., Coyne M., Dahlke C., Mays A. D., Dombroski M., Donnelly M., Ely D., Esparham S., Fosler C., Gire H., Glanowski S., Glasser K., Glodek A., Gorokhov M., Graham K., Gropman B., Harris M., Heil J., Henderson S., Hoover J., Jennings D., Jordan C., Jordan J., Kasha J., Kagan L., Kraft C., Levitsky A., Lewis M., Liu X. J., Lopez J., Ma D., Majoros W., McDaniel J., Murphy S., Newman M., Nguyen T., Nguyen N., Nodell M., Pan S., Peck J., Peterson M., Rowe W., Sanders R., Scott J., Simpson M., Smith T., Sprague A., Stockwell T., Turner R., Venter E., Wang M., Wen M. Y., Wu D., Wu M., Xia A., Zandieh A., Zhu X. H. The sequence of the human genome. *Science*. 2001, 291: 1304-+.

- (9) Arber W., Dussoix D. Host Specificity of DNA Produced by Escherichia-Coli .1. Host Controlled Modification of Bacteriophage Lambda. *Journal of Molecular Biology*. 1962, 5: 18-36.
- (10) Dussoix D., Arber W. Host Specificity of DNA Produced by Escherichia-Coli .2. Control over Acceptance of DNA from Infecting Phage Lambda. *Journal of Molecular Biology*. 1962, 5: 37-49.
- (11) Dussoix D., Arber W. Host Specificity of DNA Produced by Escherichia Coli .4. Host Specificity of Infectious DNA from Bacteriophage Lambda. *Journal of Molecular Biology*. 1965, 11: 238-246.
- (12) Smith H. O., Wilcox K. W. A Restriction Enzyme from Hemophilus-Influenzae .1. Purification and General Properties. *Journal of Molecular Biology*. 1970, 51: 379-&.
- (13) Kelly T. J., Smith H. O. A Restriction Enzyme from Hemophilus-Influenzae .2. Base Sequence of Recognition Site. *Journal of Molecular Biology*. 1970, 51: 393-&.
- (14) Roy P. H., Smith H. O. DNA methylases of Hemophilus influenzae Rd. I. Purification and properties. *Journal of Molecular Biology*. 1973, 81: 427-444.
- (15) Roy P. H., Smith H. O. DNA methylases of Hemophilus influenzae Rd. II. Partial recognition site base sequences. *Journal of Molecular Biology*. 1973, 81: 445-459.
- (16) Roberts R. J. How restriction enzymes became the workhorses of molecular biology. *Proceedings of the National Academy of Sciences of the United States of America*. 2005, 102: 5905-5908.
- (17) Sack G. H., Nathans D. Studies of Sv40 DNA .6. Cleavage of Sv40 DNA by Restriction Endonuclease from Hemophilus-Parainfluenzae. *Virology*. 1973, 51: 517-520.

- (18) Adler S. P., Nathans D. Studies of Sv 40 DNA .5. Conversion of Circular to Linear Sv 40 DNA by Restriction Endonuclease from Escherichia-Coli-B. *Biochimica Et Biophysica Acta*. 1973, 299: 177-188.
- (19) Danna K., Nathans D. Studies of Sv40 DNA .1. Specific Cleavage of Simian Virus 40 DNA by Restriction Endonuclease of Hemophilus Influenzae. *Proceedings of the National Academy of Sciences of the United States of America*. 1971, 68: 2913-&.
- (20) The Nobel Prize in Physiology or Medicine 1978: Press Release. *Nobelförsamlingen Karolinska Institutet (The Nobel Assembly at the Karolinska Institute)*. 1978: http://nobelprize.org/nobel_prizes/medicine/laureates/1978/press.html.
- (21) Kovall R. A., Matthews B. W. Type II restriction endonucleases: structural, functional and evolutionary relationships. *Current Opinion in Chemical Biology*. 1999, 3: 578-583.
- (22) Nikolajewa S., Beyer A., Friedel M., Hollunder J., Wilhelm T. Common patterns in type II restriction enzyme binding sites. *Nucleic Acids Research*. 2005, 33: 2726-2733.
- (23) Waterman M. S. Frequencies of restriction sites. *Nucleic Acids Research*. 1983, 11: 8951-8956.
- (24) Xu C. S., Kucera R. B., Roberts R. J., Guo H. C. An asymmetric complex of restriction endonuclease MspI on its palindromic DNA recognition site. *Structure*. 2004, 12: 1741-1747.
- (25) Tan C. M., Reza F., You L. C. Noise-limited frequency signal transmission in gene circuits. *Biophysical Journal*. 2007, 93: 3753-3761.
- (26) Lipan O., Wong W. H. The use of oscillatory signals in the study of genetic networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2005, 102: 7063-7068.

- (27) Gadgil C., Lee C. H., Othmer H. G. A stochastic analysis of first-order reaction networks. *Bulletin of Mathematical Biology*. 2005, 67: 901-946.
- (28) Thattai M., van Oudenaarden A. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2001, 98: 8614-8619.
- (29) Walleczek J. *Self-Organized Biological Dynamics and Nonlinear Control*. 2000. Cambridge University Press, Cambridge, UK.
- (30) Gillespie D. T. Exact Stochastic Simulation of Coupled Chemical-Reactions. *Journal of Physical Chemistry*. 1977, 81: 2340-2361.
- (31) Berman H. M. The Protein Data Bank: a historical perspective. *Acta Crystallographica Section A*. 2008, 64: 88-95.
- (32) Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E. The Protein Data Bank. *Nucleic Acids Research*. 2000, 28: 235-242.
- (33) Berman H. M., Olson W. K., Beveridge D. L., Westbrook J., Gelbin A., Demeny T., Hsieh S. H., Srinivasan A. R., Schneider B. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal*. 1992, 63: 751-759.
- (34) Roberts R. J., Vincze T., Posfai J., Macelis D. REBASE--restriction enzymes and DNA methyltransferases. *Nucleic Acids Research*. 2005, 33: D230-232.
- (35) Roberts R. J., Vincze T., Posfai J., Macelis D. REBASE - enzymes and genes for DNA restriction and modification. *Nucleic Acids Research*. 2007, 35: D269-D270.

- (36) Elhassan M. A., Calladine C. R. The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA - a New Local Calculation Scheme. *Journal of Molecular Biology*. 1995, 251: 648-664.
- (37) Lu X. J., ElHassan M. A., Hunter C. A. Structure and conformation of helical nucleic acids: Analysis program (SCHNAaP). *Journal of Molecular Biology*. 1997, 273: 668-680.
- (38) Bansal M., Bhattacharyya D., Ravi B. Nuparm and Nucgen - Software for Analysis and Generation of Sequence-Dependent Nucleic-Acid Structures. *Computer Applications in the Biosciences*. 1995, 11: 281-287.
- (39) Cornell W. D., Cieplak P., Bayly C. I., Gould I. R., Merz K. M., Ferguson D. M., Spellmeyer D. C., Fox T., Caldwell J. W., Kollman P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *Journal of the American Chemical Society*. 1996, 118: 2309-2309.
- (40) Guex N., Peitsch M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*. 1997, 18: 2714-2723.
- (41) Cheng X., Balendiran K., Schildkraut I., Anderson J. E. Structure of PvuII endonuclease with cognate DNA. *Embo Journal*. 1994, 13: 3927-3935.
- (42) Horton J. R., Bonventre J., Cheng X. D. How is modification of the DNA substrate recognized by the PvuII restriction endonuclease? *Biological Chemistry*. 1998, 379: 451-458.
- (43) Horton J. R., Nastri H. G., Riggs P. D., Cheng X. Asp34 of PvuII endonuclease is directly involved in DNA minor groove recognition and indirectly involved in catalysis. *Journal of Molecular Biology*. 1998, 284: 1491-1504.

- (44) Horton J. R., Cheng X. D. PvuII endonuclease contains two calcium ions in active sites. *Journal of Molecular Biology*. 2000, 300: 1049-1056.
- (45) Athanasiadis A., Vlassi M., Kotsifaki D., Tucker P. A., Wilson K. S., Kokkinidis M. Crystal-Structure of PvuII Endonuclease Reveals Extensive Structural Homologies to Ecorv. *Nature Structural Biology*. 1994, 1: 469-475.
- (46) Spyridaki A., Matzen C., Lanio T., Jeltsch A., Simoncsits A., Athanasiadis A., Scheuring-Vanamee E., Kokkinidis M., Pingoud A. Structural and biochemical characterization of a new Mg²⁺ binding site near Tyr94 in the restriction endonuclease PvuII. *Journal of Molecular Biology*. 2003, 331: 395-406.
- (47) Lu X. J., Olson W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*. 2003, 31: 5108-5121.
- (48) Zheng G. H., Lu X. J., Olson W. K. Web 3DNA-a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Research*. 2009, 37: W240-W246.
- (49) Lu X. J., Olson W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*. 2008, 3: 1213-1227.
- (50) DeLano W. L. In. February 6, 2006. DeLano Scientific LLC, San Carlos, CA, USA.
- (51) Kuhlman B., Baker D. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*. 2000, 97: 10383-10388.
- (52) Akutsu T. (ed.) *NP-hardness results for protein side-chain packing*. 1997.

- (53) Pierce N. A., Winfree E. Protein design is NP-hard. *Protein Engineering*. 2002, 15: 779-782.
- (54) Street A. G., Mayo S. L. Computational protein design. *Structure with Folding & Design*. 1999, 7: R105-R109.
- (55) Kraemer-Pecore C. M., Wollacott A. M., Desjarlais J. R. Computational protein design. *Current Opinion in Chemical Biology*. 2001, 5: 690-695.
- (56) Chazelle B., Kingsford C., Singh M. A semidefinite programming approach to side chain positioning with new rounding strategies. *Inform Journal on Computing*. 2004, 16: 380-392.
- (57) Poole A. M., Ranganathan R. Knowledge-based potentials in protein design. *Current Opinion in Structural Biology*. 2006, 16: 508-513.
- (58) Mirny L. A., Gelfand M. S. Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Research*. 2002, 30: 1704-1711.
- (59) Morozov A. V., Havranek J. J., Baker D., Siggia E. D. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Research*. 2005, 33: 5781-5798.
- (60) Chevalier B. S., Kortemme T., Chadsey M. S., Baker D., Monnat R. J., Stoddard B. L. Design, activity, and structure of a highly specific artificial endonuclease. *Molecular Cell*. 2002, 10: 895-905.
- (61) Fajardo-Sanchez E., Stricher F., Paques F., Isalan M., Serrano L. Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic Acids Research*. 2008, 36: 2163-2173.
- (62) Bahar I., Atilgan A. R., Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*. 1997, 2: 173-181.

- (63) Bahar I., Rader A. J. Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology*. 2005, 15: 586-592.
- (64) Eyal E., Yang L. W., Bahar I. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*. 2006, 22: 2619-2627.
- (65) Ashworth J., Havranek J. J., Duarte C. M., Sussman D., Monnat R. J., Stoddard B. L., Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*. 2006, 441: 656-659.
- (66) Thyme S. B., Jarjour J., Takeuchi R., Havranek J. J., Ashworth J., Scharenberg A. M., Stoddard B. L., Baker D. Exploitation of binding energy for catalysis and design. *Nature*. 2009, 461: 1300-U1142.
- (67) Rothlisberger D., Khersonsky O., Wollacott A. M., Jiang L., DeChancie J., Betker J., Gallaher J. L., Althoff E. A., Zanghellini A., Dym O., Albeck S., Houk K. N., Tawfik D. S., Baker D. Kemp elimination catalysts by computational enzyme design. *Nature*. 2008, 453: 190-U194.
- (68) Jiang L., Althoff E. A., Clemente F. R., Doyle L., Rothlisberger D., Zanghellini A., Gallaher J. L., Betker J. L., Tanaka F., Barbas C. F., Hilvert D., Houk K. N., Stoddard B. L., Baker D. De novo computational design of retro-aldol enzymes. *Science*. 2008, 319: 1387-1391.
- (69) Voigt C. A., Mayo S. L., Arnold F. H., Wang Z. G. Computational method to reduce the search space for directed protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2001, 98: 3778-3783.
- (70) Moore G. L., Maranas C. D. Modeling DNA mutation and recombination for directed evolution experiments. *Journal of Theoretical Biology*. 2000, 205: 483-503.

- (71) Drummond D. A., Iverson B. L., Georgiou G., Arnold F. H. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *Journal of Molecular Biology*. 2005, 350: 806-816.
- (72) Chen R. D. Enzyme engineering: rational redesign versus directed evolution. *Trends in Biotechnology*. 2001, 19: 13-14.
- (73) Collins C. H., Yokobayashi Y., Umeno D., Arnold F. H. Engineering proteins that bind, move, make and break DNA (vol 14, pg 371, 2003). *Current Opinion in Biotechnology*. 2003, 14: 665-665.
- (74) Kinney J. B., Tkacik G., Callan C. G. Precise physical models of protein - DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences of the United States of America*. 2007, 104: 501-506.
- (75) Meng X., Thibodeau-Beganny S., Jiang T., Joung J. K., Wolfe S. A. Profiling the DNA-binding specificities of engineered Cys2His2 zinc finger domains using a rapid cell-based method. *Nucleic Acids Research*. 2007, 35: -.
- (76) Bommarius A. S., Broering J. M., Chaparro-Riggers J. F., Polizzi K. M. High-throughput screening for enhanced protein stability. *Current Opinion in Biotechnology*. 2006, 17: 606-610.
- (77) Yoo T. H., Link A. J., Tirrell D. A. Evolution of a fluorinated green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*. 2007, 104: 13887-13890.
- (78) Malakauskas S. M., Mayo S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology*. 1998, 5: 470-475.
- (79) Miyazaki K., Wintrode P. L., Grayling R. A., Rubingh D. N., Arnold F. H. Directed evolution study of temperature adaptation in a psychrophilic enzyme. *Journal of Molecular Biology*. 2000, 297: 1015-1026.

- (80) Wintrobe P. L., Arnold F. H. Directed evolution of a psychrophilic enzyme. *FASEB Journal*. 1999, 13: A1535-A1535.
- (81) Arnold F. H. Directed evolution: Creating biocatalysts for the future. *Chemical Engineering Science*. 1996, 51: 5091-5102.
- (82) Arnold F. H., Volkov A. A. Directed evolution of biocatalysts. *Current Opinion in Chemical Biology*. 1999, 3: 54-59.
- (83) Arnold F. H. Engineering Enzymes for Nonnatural Environments - Improved Biocatalysts for the Chemical and Biotechnology Industries. *Protein Engineering*. 1993, 6: 59-59.
- (84) Voigt C. A., Gordon D. B., Mayo S. L. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology*. 2000, 299: 789-803.
- (85) Voigt C. A., Mayo S. L., Arnold F. H., Wang Z. G. Computationally focusing the directed evolution of proteins. *Journal of Cellular Biochemistry*. 2001: 58-63.
- (86) Nasri M., Thomas D. Alteration of the specificity of PvuII restriction endonuclease. *Nucleic Acids Research*. 1987, 15: 7677-7687.
- (87) Zaremba M., Siksnys V. Molecular scissors under light control. *Proceedings of the National Academy of Sciences of the United States of America*. 2010, 107: 1259-1260.
- (88) Schierling B., Noel A. J., Wende W., Hien L. T., Volkov E., Kubareva E., Oretskaya T., Kokkinidis M., Rompp A., Spengler B., Pingoud A. Controlling the enzymatic activity of a restriction enzyme by light. *Proceedings of the National Academy of Sciences of the United States of America*. 2010, 107: 1361-1366.

- (89) Bloom J. D., Arnold F. H. In the light of directed evolution: Pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2009, 106: 9995-10000.
- (90) Otten L. G., Hollmann F., Arends I. W. C. E. Enzyme engineering for enantioselectivity: from trial-and-error to rational design? *Trends in Biotechnology*. 2010, 28: 46-54.
- (91) Bishop A. C., Chen V. L. Brought to life: targeted activation of enzyme function with small molecules. *J Chem Biol*. 2009, 2: 1-9.
- (92) Humphrey W., Dalke A., Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*. 1996, 14: 33-&.
- (93) Pettersen E. F., Goddard T. D., Huang C. C., Couch G. S., Greenblatt D. M., Meng E. C., Ferrin T. E. UCSF chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. 2004, 25: 1605-1612.
- (94) Creighton T. E. *Proteins, Structures and Molecular Properties*. 1993, 2nd ed. W. H. Freeman and Co.
- (95) Word J. M., Lovell S. C., Richardson J. S., Richardson D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*. 1999, 285: 1735-1747.
- (96) Vriend G. What If - a Molecular Modeling and Drug Design Program. *Journal of Molecular Graphics*. 1990, 8: 52-56.
- (97) Davis I. W., Leaver-Fay A., Chen V. B., Block J. N., Kapral G. J., Wang X., Murray L. W., Arendall W. B., 3rd, Snoeyink J., Richardson J. S., Richardson D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*. 2007, 35: W375-383.

- (98) Davis I. W., Murray L. W., Richardson J. S., Richardson D. C. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research*. 2004, 32: W615-619.
- (99) Nielsen J. E., Vriend G. Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK(a) calculations. *Proteins-Structure Function and Genetics*. 2001, 43: 403-412.
- (100) Ponder J. W., Richards F. M. Tertiary Templates for Proteins - Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *Journal of Molecular Biology*. 1987, 193: 775-791.
- (101) Dunbrack R. L., Karplus M. Backbone-Dependent Rotamer Library for Proteins - Application to Side-Chain Prediction. *Journal of Molecular Biology*. 1993, 230: 543-574.
- (102) Dunbrack R. L., Cohen F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*. 1997, 6: 1661-1681.
- (103) Lovell S. C., Word J. M., Richardson J. S., Richardson D. C. The penultimate rotamer library. *Proteins*. 2000, 40: 389-408.
- (104) Lovell S. C., Davis I. W., Adrendall W. B., de Bakker P. I. W., Word J. M., Prisant M. G., Richardson J. S., Richardson D. C. Structure validation by C alpha geometry: phi,psi and C beta deviation. *Proteins-Structure Function and Genetics*. 2003, 50: 437-450.
- (105) Sauton N., Lagorce D., Villoutreix B. O., Miteva M. A. MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics*. 2008, 9: -.
- (106) Shetty R. P., de Bakker P. I. W., DePristo M. A., Blundell T. L. Advantages of fine-grained side chain conformer libraries. *Protein Engineering*. 2003, 16: 963-969.

- (107) Lassila J. K., Privett H. K., Allen B. D., Mayo S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America*. 2006, 103: 16710-16715.
- (108) Mendes J., Guerois R., Serrano L. Energy estimation in protein design. *Current Opinion in Structural Biology*. 2002, 12: 441-446.
- (109) Boas F. E., Harbury P. B. Potential energy functions for protein design. *Current Opinion in Structural Biology*. 2007, 17: 199-204.
- (110) Gordon D. B., Marshall S. A., Mayo S. L. Energy functions for protein design. *Current Opinion in Structural Biology*. 1999, 9: 509-513.
- (111) Lazaridis T., Karplus M. Effective energy function for proteins in solution. *Proteins*. 1999, 35: 133-152.
- (112) Georgiev I., Lilien R. H., Donald B. R. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *Journal of Computational Chemistry*. 2008, 29: 1527-1542.
- (113) Georgiev I. Novel algorithms for computational protein design, with applications to enzyme redesign and small-molecule inhibitor design. (Dissertation). 2009, Duke University, Durham.
- (114) Georgiev I., Lilien R. H., Donald B. R. Improved Pruning algorithms and Divide-and-Conquer strategies for Dead-End Elimination, with application to protein design. *Bioinformatics*. 2006, 22: E174-E183.
- (115) Leach A. R., Lemon A. P. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*. 1998, 33: 227-239.

- (116) Najmanovich R., Kuttner J., Sobolev V., Edelman M. Side-chain flexibility in proteins upon ligand binding. *Proteins*. 2000, 39: 261-268.
- (117) Wang T., Tomic S., Gabdoulline R. R., Wade R. C. How optimal are the binding energetics of barnase and barstar? *Biophysical Journal*. 2004, 87: 1618-1630.
- (118) Dill K. A. Dominant forces in protein folding. *Biochemistry*. 1990, 29: 7133-7155.
- (119) Dill K. A. Polymer principles and protein folding. *Protein Science*. 1999, 8: 1166-1180.
- (120) Tanimura R., Kidera A., Nakamura H. Determinants of protein side-chain packing. *Protein Science*. 1994, 3: 2358-2365.
- (121) Desjarlais J. R., Clarke N. D. Computer search algorithms in protein modification and design. *Curr Opin Struct Biol*. 1998, 8: 471-475.
- (122) Looger L. L., Hellinga H. W. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *Journal of Molecular Biology*. 2001, 307: 429-445.
- (123) Luscombe N. M., Thornton J. M. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology*. 2002, 320: 991-1009.
- (124) Lilien R. H., Stevens B. W., Anderson A. C., Donald B. R. A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *Journal of Computational Biology*. 2005, 12: 740-761.

- (125) Kingsford C. L., Chazelle B., Singh M. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*. 2005, 21: 1028-1036.
- (126) Desmet J., De Maeyer M., Lasters I. Theoretical and algorithmical optimization of the dead-end elimination theorem. *Pacific Symposium on Biocomputing*. 1997: 122-133.
- (127) Desmet J., Maeyer M. D., Hazes B., Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*. 1992, 356: 539-542.
- (128) Lasters I., De Maeyer M., Desmet J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Engineering*. 1995, 8: 815-822.
- (129) Yanover C., Fromer M., Shifman J. M. Dead-end elimination for multistate protein design. *Journal of Computational Chemistry*. 2007, 28: 2122-2129.
- (130) Gordon D. B., Mayo S. L. Branch-and Terminate: a combinatorial optimization algorithm for protein design. *Structure*. 1999, 7: 1089-1098.
- (131) Goldstein R. F. Efficient Rotamer Elimination Applied to Protein Side-Chains and Related Spin-Glasses. *Biophysical Journal*. 1994, 66: 1335-1340.
- (132) Pierce N. A., Spriet J. A., Desmet J., Mayo S. L. Conformational splitting: A more powerful criterion for dead-end elimination. *Journal of Computational Chemistry*. 2000, 21: 999-1009.
- (133) Georgiev I., Lilien R. H., Donald B. R. Improved Pruning algorithms and Divide-and-Conquer strategies for Dead-End Elimination, with application to protein design. *Bioinformatics*. 2006, 22: e174-183.

- (134) Lasters I., Desmet J., DeMaeyer M. Dead-end based modeling tools to explore the sequence space that is compatible with a given scaffold. *Journal of Protein Chemistry*. 1997, 16: 449-452.
- (135) Gordon D. B., Hom G. K., Mayo S. L., Pierce N. A. Exact rotamer optimization for protein design. *Journal of Computational Chemistry*. 2003, 24: 232-243.
- (136) Brakmann S. Discovery of superior enzymes by directed molecular evolution. *Chembiochem*. 2001, 2: 865-871.
- (137) Hellinga H. W., Richards F. M. Optimal sequence selection in proteins of known structure by simulated evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 1994, 91: 5803-5807.
- (138) Arnold F. H., Volkov A. A. Directed evolution of biocatalysts. *Current Opinion in Chemical Biology*. 1999, 3: 54-59.
- (139) Voigt C. A., Martinez C., Wang Z. G., Mayo S. L., Arnold F. H. Protein building blocks preserved by recombination. *Nature Structural Biology*. 2002, 9: 553-558.
- (140) Song J. K., Chung B., Oh Y. H., Rhee J. S. Construction of DNA-shuffled and incrementally truncated libraries by a mutagenic and unidirectional reassembly method: changing from a substrate specificity of phospholipase to that of lipase. *Applied and Environmental Microbiology*. 2002, 68: 6146-6151.
- (141) Ostermeier M., Nixon A. E., Benkovic S. J. Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorganic & Medicinal Chemistry*. 1999, 7: 2139-2144.
- (142) Cochran J. R., Kim Y. S., Lippow S. M., Rao B., Wittrup K. D. Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Engineering, Design and Selection*. 2006, 19: 245-253.

- (143) Pingoud A., Fuxreiter M., Pingoud V., Wende W. Type II restriction endonucleases: structure and mechanism. *Cellular and Molecular Life Sciences*. 2005, 62: 685-707.
- (144) Pingoud A. *Restriction endonucleases*. 2004. Springer, Berlin ; New York.
- (145) Adams G. M., Blumenthal R. M. The PvuII DNA (cytosine-N4)-methyltransferase comprises two trypsin-defined domains, each of which binds a molecule of S-adenosyl-L-methionine. *Biochemistry*. 1997, 36: 8284-8292.
- (146) Gong W. M., OGara M., Blumenthal R. M., Cheng X. D. Structure of PvuII DNA (cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. *Nucleic Acids Research*. 1997, 25: 2702-2715.
- (147) Kim Y. C., Grable J. C., Love R., Greene P. J., Rosenberg J. M. Refinement of EcoRI endonuclease crystal structure: a revised protein chain tracing. *Science*. 1990, 249: 1307-1309.
- (148) Schottler S., Wenz C., Lanio T., Jeltsch A., Pingoud A. Protein engineering of the restriction endonuclease EcoRV--structure-guided design of enzyme variants that recognize the base pairs flanking the recognition site. *European Journal of Biochemistry*. 1998, 258: 184-191.
- (149) Winkler F. K., Banner D. W., Oefner C., Tsernoglou D., Brown R. S., Heathman S. P., Bryan R. K., Martin P. D., Petratos K., Wilson K. S. The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO Journal*. 1993, 12: 1781-1795.
- (150) Lanio T., Selent U., Wenz C., Wende W., Schulz A., Adiraj M., Katti S. B., Pingoud A. EcoRV-T94V: A mutant restriction endonuclease with an altered substrate specificity towards modified oligodeoxynucleotides. *Protein Engineering*. 1996, 9: 1005-1010.

- (151) Newman M., Strzelecka T., Dorner L. F., Schildkraut I., Aggarwal A. K. Structure of restriction endonuclease bamhi phased at 1.95 Å resolution by MAD analysis. *Structure*. 1994, 2: 439-452.
- (152) Roberts R. J. Restriction enzymes and their isoschizomers. *Nucleic Acids Research*. 1989, 17 Suppl: r347-387.
- (153) Thompson J. D., Higgins D. G., Gibson T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. 1994, 22: 4673-4680.
- (154) Murzin A. G., Brenner S. E., Hubbard T., Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*. 1995, 247: 536-540.
- (155) Kyte J., Doolittle R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*. 1982, 157: 105-132.
- (156) Duan Y., Wu C., Chowdhury S., Lee M. C., Xiong G., Zhang W., Yang R., Cieplak P., Luo R., Lee T., Caldwell J., Wang J., Kollman P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*. 2003, 24: 1999-2012.
- (157) Bansal M., Bhattacharyya D., Ravi B. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Computer Applications in the Biosciences*. 1995, 11: 281-287.
- (158) Tian J., Gong H., Sheng N., Zhou X., Gulari E., Gao X., Church G. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*. 2004, 432: 1050-1054.

- (159) Reza F., Zuo P., Tian J. Protein Interfacial Pocket Engineering via Coupled Computational Filtering and Biological Focusing Criterion. *Annals of Biomedical Engineering. Special Issue: Systems Biology, Bioinformatics, and Computational Biology*. 2007, 35: 1026-1036.
- (160) Balendiran K., Bonventre J., Knott R., Jack W., Benner J., Schildkraut I., Anderson J. E. Expression, purification, and crystallization of restriction endonuclease PvuII with DNA containing its recognition site. *Proteins*. 1994, 19: 77-79.
- (161) Jones S., van Heyningen P., Berman H. M., Thornton J. M. Protein-DNA interactions: A structural analysis. *Journal of Molecular Biology*. 1999, 287: 877-896.
- (162) Doyle K., Zhang Y., Baer R., Bina M. Distinguishable patterns of protein-DNA interactions involving complexes of basic helix-loop-helix proteins. *Journal of Biological Chemistry*. 1994, 269: 12099-12105.
- (163) Deremble C., Lavery R. Macromolecular recognition. *Current Opinion in Structural Biology*. 2005, 15: 171-175.
- (164) Donald J. E., Chen W. W., Shakhnovich E. I. Energetics of protein-DNA interactions. *Nucleic Acids Research*. 2007, 35: 1039-1047.
- (165) Endres R. G., Schulthess T. C., Wingreen N. S. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*. 2004, 57: 262-268.
- (166) Giudice E., Lavery R. Simulations of nucleic acids and their complexes. *Accounts of Chemical Research*. 2002, 35: 350-357.
- (167) Harrison S. C., Aggarwal A. K. DNA recognition by proteins with the helix-turn-helix motif. *Annual Review of Biochemistry*. 1990, 59: 933-969.

- (168) Luscombe N. M., Austin S. E., Berman H. M., Thornton J. M. An overview of the structures of protein-DNA complexes. *Genome Biology*. 2000, 1: REVIEWS001.
- (169) Steffen N. R., Murphy S. D., Toller L., Hatfield G. W., Lathrop R. H. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics*. 2002, 18 Suppl 1: S22-30.
- (170) Chen Z. L., Wen F., Sun N., Zhao H. M. Directed evolution of homing endonuclease I-SceI with altered sequence specificity. *Protein Engineering Design & Selection*. 2009, 22: 249-256.
- (171) Doyon J. B., Pattanayak V., Meyer C. B., Liu D. R. Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *Journal of the American Chemical Society*. 2006, 128: 2477-2484.
- (172) Lanio T., Jeltsch A., Pingoud A. Towards the design of rare cutting restriction endonucleases: using directed evolution to generate variants of EcoRV differing in their substrate specificity by two orders of magnitude. *Journal of Molecular Biology*. 1998, 283: 59-69.
- (173) Morgan R. D., Luyten Y. A. Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. *Nucleic Acids Research*. 2009.
- (174) Lippow S. M., Aha P. M., Parker M. H., Blake W. J., Baynes B. M., Lipovsek D. Creation of a type IIS restriction endonuclease with a long recognition sequence. *Nucleic Acids Research*. 2009, 37: 3061-3073.
- (175) Eisenschmidt K., Lanio T., Simoncsits A., Jeltsch A., Pingoud V., Wende W., Pingoud A. Developing a programmed restriction endonuclease for highly specific DNA cleavage. *Nucleic Acids Research*. 2005, 33: 7039-7047.
- (176) Chin J. Y., Kuan J. Y., Lonkar P. S., Krause D. S., Seidman M. M., Peterson K. R., Nielsen P. E., Kole R., Glazer P. M. Correction of a splice-site mutation in the beta-globin gene stimulated by triplex-

forming peptide nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America*. 2008, 105: 13514-13519.

- (177) Alvizo O., Allen B. D., Mayo S. L. Computational protein design promises to revolutionize protein engineering. *Biotechniques*. 2007, 42: 31-38.
- (178) Goodsell D. S., Olson A. J. Structural symmetry and protein function. *Annual Reviews of Biophysics and Biomolecular Structure*. 2000, 29: 105-153.
- (179) Tao T., Blumenthal R. M. Sequence and Characterization of PvuIIr, the PvuII Endonuclease Gene, and of PvuIIc, Its Regulatory Gene. *Journal of Bacteriology*. 1992, 174: 3395-3398.
- (180) Rice M. R., Koons M. D., Blumenthal R. M. Substrate recognition by the PvuII endonuclease: binding and cleavage of CAG(5m)CTG sites. *Nucleic Acids Research*. 1999, 27: 1032-1038.
- (181) Chen C. Y., Georgiev I., Anderson A. C., Donald B. R. Computational structure-based redesign of enzyme activity. *Proceedings of the National Academy of Sciences of the United States of America*. 2009, 106: 3764-3769.
- (182) Dongarra J., Walker D., Lusk E., Knighten B., Snir M., Geist A., Otto S., Hempel R., Lusk E., Gropp W., Cownie J., Skjellum T., Clarke L., Littlefield R., Sears M., Husslederman S., Anderson E., Berryman S., Feeney J., Frye D., Hart L., Ho A., Kohl J., Madams P., Mosher C., Pierce P., Schikuta E., Voigt R. G., Babb R., Bjornson R., Fernando V., Glendinning I., Haupt T., Ho C. T. H., Krauss S., Mainwaring A., Nessett D., Ranka S., Singh A., Weeks D., Baron J., Doss N., Fineberg S., Greenberg A., Heller D., Howell G., Leary B., Mcbryan O., Pacheco P., Rigsbee P., Sussman A., Wheat S., Barszcz E., Elster A., Flower J., Harrison R., Henderson T., Kapenga J., Maccabe A., Mckinley P., Palmer H., Robison A., Tomlinson R., Zenith S. Special Issue - Mpi - a Message-Passing Interface Standard. *International Journal of Supercomputer Applications and High Performance Computing*. 1994, 8: 165-&.

- (183) Walker D. W., Dongarra J. J. MPI: A standard message passing interface. *Supercomputer*. 1996, 12: 56-68.
- (184) Georgiev I., Donald B. R. OSPREY User Manual v 1.0. <http://www.cs.duke.edu/donaldlab/software/osprey/OSPREY.pdf>. 2009.
- (185) Lilien R. H., Stevens B. W., Anderson A. C., Donald B. R. A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *J Comput Biol*. 2005, 12: 740-761.
- (186) Nastri H. G., Evans P. D., Walker I. H., Riggs P. D. Catalytic and DNA binding properties of PvuII restriction endonuclease mutants. *Journal of Biological Chemistry*. 1997, 272: 25761-25767.
- (187) Kuruma Y., Nishiyama K., Shimizu Y., Muller M., Ueda T. Development of a minimal cell-free translation system for the synthesis of presecretory and integral membrane proteins. *Biotechnology Progress*. 2005, 21: 1243-1251.
- (188) Nirenberg M., Matthaei J. H. Dependence of Cell-Free Protein Synthesis in E Coli Upon Naturally Occurring or Synthetic Polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*. 1961, 47: 1588-1602.
- (189) Shimizu Y., Inoue A., Tomari Y., Suzuki T., Yokogawa T., Nishikawa K., Ueda T. Cell-free translation reconstituted with purified components. *Nature Biotechnology*. 2001, 19: 751-755.
- (190) Shimizu Y., Kanamori T., Ueda T. Protein synthesis by pure translation systems. *Methods*. 2005, 36: 299-304.
- (191) Swartz J. A PURE approach to constructive biology. *Nature Biotechnology*. 2001, 19: 732-733.

- (192) Chong L. T., Swope W. C., Pitera J. W., Pande V. S. Kinetic computational alanine scanning: Application to p53 oligomerization. *Journal of Molecular Biology*. 2006, 357: 1039-1049.
- (193) Kortemme T., Kim D. E., Baker D. Computational alanine scanning of protein-protein interfaces. *Science STKE*. 2004, 2004: pl2.
- (194) Massova I., Kollman P. A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society*. 1999, 121: 8133-8143.
- (195) Huo S., Massova I., Kollman P. A. Computational alanine scanning of the 1 : 1 human growth hormone-receptor complex. *Journal of Computational Chemistry*. 2002, 23: 15-27.
- (196) Altman M. D., Ali A., Reddy G. S. K. K., Nalam M. N. L., Anjum S. G., Cao H., Chellappan S., Kairys V., Fernandes M. X., Gilson M. K., Schiffer C. A., Rana T. M., Tidor B. HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *Journal of the American Chemical Society*. 2008, 130: 6099-6113.
- (197) Suarez M., Jaramillo A. Challenges in the computational design of proteins. *Journal of the Royal Society Interface*. 2009, 6: -.
- (198) Busch M. S. A., Mignon D., Simonson T. Computational protein design as a tool for fold recognition. *Proteins-Structure Function and Bioinformatics*. 2009, 77: 139-158.
- (199) Busch M. S. A., Lopes A., Mignon D., Simonson T. Computational protein design: Software implementation, parameter optimization, and performance of a simple model. *Journal of Computational Chemistry*. 2008, 29: 1092-1102.

- (200) Chowdry A., Reynolds K., Voorhies M., Pokala N., Handel T. An object-oriented library for computational protein design. *Biophysical Journal*. 2005, 88: 332a-332a.
- (201) Chowdry A. B., Reynolds K. A., Hanes M. S., Voorhies M., Pokala N., Handel T. M. Software news and update an object-oriented library for computational protein design. *Journal of Computational Chemistry*. 2007, 28: 2378-2388.
- (202) Liu Y., Kuhlman B. RosettaDesign server for protein design. *Nucleic Acids Research*. 2006, 34: W235-W238.
- (203) Dahiyat B. I., Mayo S. L. De novo protein design: Fully automated sequence selection. *Science*. 1997, 278: 82-87.
- (204) Loksha I. V., Maiolo J. R., Hong C. W., Ng A., Snow C. D. SHARPEN-Systematic Hierarchical Algorithms for Rotamers and Proteins on an Extended Network. *Journal of Computational Chemistry*. 2009, 30: 999-1005.
- (205) Dunbrack R. L. Rotamer libraries in the 21(st) century. *Current Opinion in Structural Biology*. 2002, 12: 431-440.
- (206) Mendes J., Baptista A. M., Carrondo M. A., Soares C. M. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins-Structure Function and Genetics*. 1999, 37: 530-543.
- (207) Keating A. E., Malashkevich V. N., Tidor B., Kim P. S. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proceedings of the National Academy of Sciences of the United States of America*. 2001, 98: 14825-14830.
- (208) Dahiyat B. I., Mayo S. L. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America*. 1997, 94: 10172-10177.

- (209) Georgiev I., Donald B. R. Dead-end elimination with backbone flexibility. *Bioinformatics*. 2007, 23: I185-I194.
- (210) Hong E. J., Lippow S. M., Tidor B., Lozano-Perez T. Rotamer Optimization for Protein Design through MAP Estimation and Problem-Size Reduction. *Journal of Computational Chemistry*. 2009, 30: 1923-1945.
- (211) Hubscher U., Maga G., Spadari S. Eukaryotic DNA polymerases. *Annual Review of Biochemistry*. 2002, 71: 133-163.
- (212) Goodman M. F., Tippin B. The expanding polymerase universe. *Nature Reviews Molecular Cell Biology*. 2000, 1: 101-109.
- (213) Friedberg E. C., Feaver W. J., Gerlach V. L. The many faces of DNA polymerases: Strategies for mutagenesis and for mutational avoidance. *Proceedings of the National Academy of Sciences of the United States of America*. 2000, 97: 5681-5683.
- (214) Patel P. H., Suzuki M., Adman E., Shinkai A., Loeb L. A. Prokaryotic DNA polymerase I: Evolution, structure, and "base flipping" mechanism for nucleotide selection. *Journal of Molecular Biology*. 2001, 308: 823-837.
- (215) Keohavong P., Thilly W. G. Fidelity of DNA-Polymerases in DNA Amplification. *Proceedings of the National Academy of Sciences of the United States of America*. 1989, 86: 9253-9257.
- (216) Gibbs R. A. DNA Amplification by the Polymerase Chain-Reaction. *Analytical Chemistry*. 1990, 62: 1202-1214.
- (217) Arnheim N., Erlich H. Polymerase Chain-Reaction Strategy. *Annual Review of Biochemistry*. 1992, 61: 131-156.

- (218) Eckert K. A., Kunkel T. A. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods and Applications*. 1991, 1: 17-24.
- (219) Young R. A. RNA Polymerase-II. *Annual Review of Biochemistry*. 1991, 60: 689-715.
- (220) Burgess R. R. RNA Polymerase. *Annual Review of Biochemistry*. 1971, 40: 711-&.
- (221) Zenkin N., Severinov K. RNA polymerase - The third class of primases. *Cellular and Molecular Life Sciences*. 2008, 65: 2280-2288.
- (222) Mullard A. Reverse transcription - Do the flip. *Nature Reviews Molecular Cell Biology*. 2008, 9: 501-501.
- (223) Temin H. M. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *Proceedings of the National Academy of Sciences of the United States of America*. 1993, 90: 6900-6903.
- (224) Preston B. D., Poiesz B. J., Loeb L. A. Fidelity of Hiv-1 Reverse-Transcriptase. *Science*. 1988, 242: 1168-1171.
- (225) Preston B. D. Reverse transcriptase fidelity and HIV-1 variation. *Science*. 1997, 275: 228-229.
- (226) Skalka A. M., Goff S. *Reverse transcriptase*. 1993. Cold Spring Harbor Laboratory Press, Plainview, N.Y.
- (227) Chen X., Pascal J., Vijayakumar S., Wilson G. M., Ellenberger T., Tomkinson A. E. Human DNA ligases I, III, and IV purification and new specific assays for these enzymes. *DNA Repair, Part B*. 2006, 409: 39-52.

- (228) Tomkinson A. E., Vijayakumar S., Pascal J. M., Ellenberger T. DNA ligases: Structure, reaction mechanism, and function. *Chemical Reviews*. 2006, 106: 687-699.
- (229) Lindahl T., Barnes D. E. Mammalian DNA Ligases. *Annual Review of Biochemistry*. 1992, 61: 251-281.
- (230) Zhou B. B. S., Elledge S. J. The DNA damage response: putting checkpoints in perspective. *Nature*. 2000, 408: 433-439.
- (231) Barzilai A., Yamamoto K. I. DNA damage responses to oxidative stress. *DNA Repair*. 2004, 3: 1109-1115.
- (232) Cheng X. D., Blumenthal R. M. Mammalian DNA methyltransferases: A structural perspective. *Structure*. 2008, 16: 341-350.
- (233) Siedlecki P., Zielenkiewicz P. Mammalian DNA methyltransferases. *Acta Biochimica Polonica*. 2006, 53: 245-256.
- (234) Cheng X. D. Structure and Function of DNA Methyltransferases. *Annual Review of Biophysics and Biomolecular Structure*. 1995, 24: 293-318.
- (235) Doerfler W. DNA methylation and gene activity. *Annu Rev Biochem*. 1983, 52: 93-124.
- (236) Bernstein E., Caudy A. A., Hammond S. M., Hannon G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*. 2001, 409: 363-366.
- (237) Okamura K., Ishizuka A., Siomi H., Siomi M. C. Distinct roles for argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes & Development*. 2004, 18: 1655-1666.

- (238) Lee Y. S., Nakahara K., Pham J. W., Kim K., He Z. Y., Sontheimer E. J., Carthew R. W. Distinct roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*. 2004, 117: 69-81.
- (239) Wool I. G. Structure and Function of Eukaryotic Ribosomes. *Annual Review of Biochemistry*. 1979, 48: 719-754.
- (240) Nissen P., Hansen J., Ban N., Moore P. B., Steitz T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science*. 2000, 289: 920-930.
- (241) Steitz T. A., Moore P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends in Biochemical Sciences*. 2003, 28: 411-418.
- (242) Wilson K. S., Noller H. F. Molecular movement inside the translational engine. *Cell*. 1998, 92: 337-349.
- (243) Schmeing T. M., Ramakrishnan V. What recent ribosome structures have revealed about the mechanism of translation. *Nature*. 2009, 461: 1234-1242.
- (244) Aggarwal A. K. Structure and Function of Restriction Endonucleases. *Current Opinion in Structural Biology*. 1995, 5: 11-19.
- (245) Pingoud A., Jeltsch A. Structure and function of type II restriction endonucleases. *Nucleic Acids Research*. 2001, 29: 3705-3727.
- (246) Bickle T. A., Kruger D. H. Biology of DNA Restriction. *Microbiological Reviews*. 1993, 57: 434-450.
- (247) Kobayashi I., Nobusato A., Kobayashi-Takahashi N., Uchiyama I. Shaping the genome - restriction-modification systems as mobile genetic elements. *Current Opinion in Genetics & Development*. 1999, 9: 649-656.

- (248) Tian J., Ma K., Saaem I. Advancing high-throughput gene synthesis technology. *Molecular BioSystems*. 2009, 5: 714-722.
- (249) Ma K. S., Reza F., Saaem I., Tian J. D. Versatile surface functionalization of cyclic olefin copolymer (COC) with sputtered SiO₂ thin film for potential BioMEMS applications. *Journal of Materials Chemistry*. 2009, 19: 7914-7920.
- (250) Smith H. O., Hutchison C. A., Pfannkoch C., Venter J. C. Generating a synthetic genome by whole genome assembly: phi X174 bacteriophage from synthetic oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*. 2003, 100: 15440-15445.
- (251) Gibson D. G., Benders G. A., Andrews-Pfannkoch C., Denisova E. A., Baden-Tillson H., Zaveri J., Stockwell T. B., Brownley A., Thomas D. W., Algire M. A., Merryman C., Young L., Noskov V. N., Glass J. I., Venter J. C., Hutchison C. A., Smith H. O. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science*. 2008, 319: 1215-1220.
- (252) Park N., Kahn J. S., Rice E. J., Hartman M. R., Funabashi H., Xu J. F., Um S. H., Luo D. High-yield cell-free protein production from P-gel. *Nature Protocols*. 2009, 4: 1759-1770.
- (253) Zhu Z., Blanchard A., Xu S.-Y., Guan S., Wei H., Zhang P., Sun D., Chan S.-H. High fidelity restriction endonucleases. 2008, patent no. 12/172,963.
- (254) Naito T., Kusano K., Kobayashi I. Selfish Behavior of Restriction-Modification Systems. *Science*. 1995, 267: 897-899.
- (255) Handa N., Nakayama Y., Sadykov M., Kobayashi I. Experimental genome evolution: large-scale genome rearrangements associated with resistance to replacement of a chromosomal restriction-modification gene complex. *Molecular Microbiology*. 2001, 40: 932-940.

- (256) Kobayashi I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research*. 2001, 29: 3742-3756.
- (257) Jeltsch A., Pingoud A. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *Journal of Molecular Evolution*. 1996, 42: 91-96.
- (258) Mrazek J., Karlin S. Detecting alien genes in bacterial genomes. *Molecular Strategies in Biological Evolution*. 1999, 870: 314-329.
- (259) Pingoud A. *Restriction endonucleases*. 2004. Springer-Verlag, Berlin ; New York.
- (260) Kobayashi I. Selfishness and death: raison d'etre of restriction, recombination and mitochondria. *Trends in Genetics*. 1998, 14: 368-374.
- (261) Kobayashi H., Kaern M., Araki M., Chung K., Gardner T. S., Cantor C. R., Collins J. J. Programmable cells: Interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2004, 101: 8414-8419.
- (262) Szymkowski D. E. Creating the next generation of protein therapeutics through rational drug design. *Current Opinion in Drug Discovery & Development*. 2005, 8: 590-600.
- (263) Farokhzad O. C., Langer R. Nanomedicine: developing smarter therapeutic and diagnostic modalities. *Advanced Drug Delivery Reviews*. 2006, 58: 1456-1459.
- (264) Nie S., Xing Y., Kim G. J., Simons J. W. Nanotechnology applications in cancer. *Annual Review of Biomedical Engineering*. 2007, 9: 257-288.
- (265) Ferrari M. Cancer nanotechnology: Opportunities and challenges. *Nature Reviews Cancer*. 2005, 5: 161-171.

- (266) Adams G. P., Weiner L. M. Monoclonal antibody therapy of cancer. *Nature Biotechnology*. 2005, 23: 1147-1157.
- (267) von Mehren M., Adams G. P., Weiner L. M. Monoclonal antibody therapy for cancer. *Annual Review of Medicine-Selected Topics in the Clinical Sciences*. 2003, 54: 343-369.
- (268) Green M. C., Murray J. L., Hortobagyi G. N. Monoclonal antibody therapy for solid tumors. *Cancer Treatment Reviews*. 2000, 26: 269-286.
- (269) Casadevall A., Dadachova E., Pirofski L. Passive antibody therapy for infectious diseases. *Nature Reviews Microbiology*. 2004, 2: 695-703.
- (270) Harley C. B. Telomerase and cancer therapeutics. *Nature Reviews Cancer*. 2008, 8: 167-179.
- (271) Shay J. W., Keith W. N. Targeting telomerase for cancer therapeutics. *British Journal of Cancer*. 2008, 98: 677-683.
- (272) Shay J. W., Wright W. E. Telomerase therapeutics for cancer: challenges and new directions. *Nature Reviews Drug Discovery*. 2006, 5: 577-584.
- (273) Kelland L. R. Overcoming the immortality of tumour cells by telomere and telomerase based cancer therapeutics - current status and future prospects. *European Journal of Cancer*. 2005, 41: 971-979.
- (274) Guittat L., Alberti P., Gomez D., De Cian A., Pennarun G., Lemarteleur T., Belmokhtar C., Paterski R., Morjani H., Trentesaux C., Mandine E., Boussin F., Mailliet P., Lacroix L., Riou J. F., Mergny J. L. Targeting human telomerase for cancer therapeutics. *Cytotechnology*. 2004, 45: 75-90.
- (275) Shay J. W., Wright W. E. Telomerase: A target for cancer therapeutics. *Cancer Cell*. 2002, 2: 257-265.

- (276) Ouellette M. M., Lee K. Telomerase: diagnostics, cancer therapeutics and tissue engineering. *Drug Discovery Today*. 2001, 6: 1231-1237.
- (277) Cunningham A. P., Love W. K., Zhang R. W., Andrews L. G., Tollefsbol T. O. Telomerase inhibition in cancer therapeutics: Approaches molecular-based. *Current Medicinal Chemistry*. 2006, 13: 2875-2888.
- (278) McCafferty J., Glover D. R. Engineering therapeutic proteins. *Current Opinion in Structural Biology*. 2000, 10: 417-420.
- (279) Lazar G. A., Marshall S. A., Plecs J. J., Mayo S. L., Desjarlais J. R. Designing proteins for therapeutic applications. *Current Opinion in Structural Biology*. 2003, 13: 513-518.
- (280) Marshall S. A., Lazar G. A., Chirino A. J., Desjarlais J. R. Rational design and engineering of therapeutic proteins. *Drug Discovery Today*. 2003, 8: 212-221.
- (281) Lippow S. M., Wittrup K. D., Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature Biotechnology*. 2007, 25: 1171-1176.
- (282) Koder R. L., Anderson J. L. R., Solomon L. A., Reddy K. S., Moser C. C., Dutton P. L. Design and engineering of an O-2 transport protein. *Nature*. 2009, 458: 305-U364.
- (283) Hefferin M. L., Tomkinson A. E. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair*. 2005, 4: 639-648.
- (284) Wilson T. E., Grawunder U., Lieber M. R. Yeast DNA ligase IV mediates non-homologous DNA end joining. *Nature*. 1997, 388: 495-498.

- (285) Knauert M. P., Glazer P. M. Triplex forming oligonucleotides: sequence-specific tools for gene targeting. *Human Molecular Genetics*. 2001, 10: 2243-2251.
- (286) Casey B. P., Glazer P. M. Gene targeting via triple-helix formation. *Progress in Nucleic Acid Research and Molecular Biology, Vol 67*. 2001, 67: 163-192.
- (287) Chin J. Y., Schleifman E. B., Glazer P. M. Repair and recombination induced by triple helix DNA. *Frontiers in Bioscience*. 2007, 12: 4288-4297.
- (288) Kim K. H., Nielsen P. E., Glazer P. M. Site-directed gene mutation at mixed sequence targets by psoralen-conjugated pseudo-complementary peptide nucleic acids. *Nucleic Acids Research*. 2007, 35: 7604-7613.
- (289) Sechler J. M. Use of restriction endonucleases against viruses, including HIV. U.S.A. Department of Health and Human Services. 1996, U.S.A. patent no. 5523232.
- (290) Carlson R. The pace and proliferation of biological technologies. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*. 2003, 1: 203-214.
- (291) Reza F., Chandran K., Feltz M., Heinz A., Josephs E., O'Brien P., Dyke B. V., Chung H., Indurkha S., Lakhani N., Lee J., Lin S., Tang N., LaBean T., You L., Yuan F., Tian J. Engineering novel synthetic biological systems. *IET Synthetic Biology*. 2007, 1: 48-52.
- (292) Levskaya A., Chevalier A. A., Tabor J. J., Simpson Z. B., Lavery L. A., Levy M., Davidson E. A., Scouras A., Ellington A. D., Marcotte E. M., Voigt C. A. Engineering *Escherichia coli* to see light - These smart bacteria 'photograph' a light pattern as a high-definition chemical image. *Nature*. 2005, 438: 441-442.
- (293) Endy D. Foundations for engineering biology. *Nature*. 2005, 438: 449-453.

- (294) Kumar S., Rai A. Synthetic biology: The intellectual property puzzle. *Texas Law Review*. 2007, 85: 1745-1768.
- (295) Rai A., Boyle J. Synthetic biology: Caught between property rights, the public domain, and the commons. *PLoS Biology*. 2007, 5: 389-393.
- (296) Angrist M., Cook-Deegan R. M. Who owns the genome? *New Atlantis*. 2006, 11: 87-96.
- (297) Chandrasekharan S., Kumar S., Valley C. M., Rai A. Proprietary science, open science and the role of patent disclosure: the case of zinc-finger proteins. *Nature Biotechnology*. 2009, 27: 140-144.
- (298) Rai A. K., Reichman J. H., Uhlir P. F., Crossman C. Pathways across the valley of death: novel intellectual property strategies for accelerated drug discovery. *Yale Journal of Health Policy, Law, and Ethics*. 2008, 8: 1-36.
- (299) Feynman R. P. *There's plenty of room at the bottom*. 1959, Lecture at California Institute of Technology (CalTech).
- (300) Drexler K. E. *Engines of creation*. 1990, New York.
- (301) Drexler K. E. Molecular Engineering - an Approach to the Development of General Capabilities for Molecular Manipulation. *Proceedings of the National Academy of Sciences of the United States of America-Physical Sciences*. 1981, 78: 5275-5278.
- (302) Asimov I., Kleiner H., Klement O. *Fantastic voyage*. 1966. Houghton Mifflin, Boston.
- (303) Asimov I. *Fantastic voyage II: destination brain*. 1987, 1st ed. Doubleday, New York.
- (304) Clarke A. C. *The collected stories of Arthur C. Clarke*. 2001, New York.

- (305) Heinlein R. A. *Waldo, and Magic Inc.* 1950. Doubleday, New York.
- (306) Crichton M. *Prey.* 2002. Harper Collins Publishers, New York.
- (307) Ludlum R., Larkin P. *The Lazarus vendetta.* 2004. St. Martin's Griffin, New York.
- (308) Leskov N. S. *Levsha.* 1944. International University Press, New York.
- (309) Lukacs C. M., Kucera R., Schildkraut R., Aggarwal A. K. Understanding the immutability of restriction enzymes: crystal structure of Bg/II and its DNA substrate at 1.5 angstrom resolution. *Nature Structural Biology.* 2000, 7: 134-140.
- (310) Conlan L. H., Dupureur C. M. Dissecting the metal ion dependence of DNA binding by PvuII endonuclease. *Biochemistry.* 2002, 41: 1335-1342.
- (311) Conlan L. H., Dupureur C. M. Multiple metal ions drive DNA association by PvuII endonuclease. *Biochemistry.* 2002, 41: 14848-14855.
- (312) Dupureur C. M., Conlan L. H. A catalytically deficient active site variant of PvuII endonuclease binds Mg(II) ions. *Biochemistry.* 2000, 39: 10921-10927.
- (313) Fromer M., Yanover C. Accurate prediction for atomic-level protein design and its application in diversifying the near-optimal sequence space. *Proteins-Structure Function and Bioinformatics.* 2009, 75: 682-705.
- (314) Lippow S. M., Tidor B. Progress in computational protein design. *Current Opinion in Biotechnology.* 2007, 18: 305-311.

- (315) Moore G. E. Cramming more components onto integrated circuits (Reprinted from *Electronics*, pg 114-117, April 19, 1965). *Proceedings of the IEEE*. 1998, 86: 82-85.
- (316) Beberg A. L., Ensign D. L., Jayachandran G., Khaliq S., Pande V. S. Folding@home: Lessons From Eight Years of Volunteer Distributed Computing. *2009 IEEE International Symposium on Parallel & Distributed Processing, Vols 1-5*. 2009: 1624-1631.
- (317) Pande V. S. Folding@Home: Using desktop grid computing to overcome fundamental barriers in biomolecular simulation. *Abstracts of Papers of the American Chemical Society*. 2005, 230: U1295-U1295.
- (318) Shirts M. R., Jayachandran G., Snow C. D., Pande V. S. Directly calculated ligand binding free energies using folding@home. *Abstracts of Papers of the American Chemical Society*. 2004, 228: U532-U532.
- (319) Pande V. S. Folding@home: Can a grid of 100,000 CPUs tackle fundamental barriers in molecular simulation? *Abstracts of Papers of the American Chemical Society*. 2004, 228: U532-U532.
- (320) Shirts M. R., Snow C. D., Pande V. S. Directly calculated ligand binding free energies using Folding@Home. *Abstracts of Papers of the American Chemical Society*. 2004, 227: U899-U900.
- (321) Pande V. S. Folding@home: Can non-equilibrium statistical mechanics and 100,000 cpus simulate protein folding in atomic detail on the millisecond timescale? *Abstracts of Papers of the American Chemical Society*. 2003, 226: U424-U424.
- (322) Das R., Qian B., Raman S., Vernon R., Thompson J., Bradley P., Khare S., Tyka M. D., Bhat D., Chivian D., Kim D. E., Sheffler W. H., Malmstrom L., Wollacott A. M., Wang C., Andre I., Baker D. Structure prediction for CABP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins-Structure Function and Bioinformatics*. 2007, 69: 118-128.

- (323) Cumbaa C. A., Jurisica I. Protein crystallization analysis on the World Community Grid. *Journal of Structural and Functional Genomics*. 2010.
- (324) Forster A. C., Church G. M. Synthetic biology projects in vitro. *Genome Research*. 2007, 17: 1-6.
- (325) Church G. M. From systems biology to synthetic biology. *Molecular Systems Biology*. 2005, 1: 2005 0032.
- (326) Carothers J. M., Goler J. A., Keasling J. D. Chemical synthesis using synthetic biology. *Current Opinion in Biotechnology*. 2009, 20: 498-503.
- (327) Lee S. K., Chou H., Ham T. S., Lee T. S., Keasling J. D. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology*. 2008, 19: 556-563.
- (328) Keasling J. D. Synthetic biology for synthetic chemistry. *ACS Chemical Biology* 2008, 3: 64-76.
- (329) Baker D., Group B. F., Church G., Collins J., Endy D., Jacobson J., Keasling J., Modrich P., Smolke C., Weiss R. Engineering life: Building a fab for biology. *Scientific American*. 2006, 294: 44-51.
- (330) Endy D. Foundations for engineering biology. *Nature*. 2005, 438: 449-453.
- (331) Hartwell L. H., Hopfield J. J., Leibler S., Murray A. W. From molecular to modular cell biology. *Nature*. 1999, 402: C47-52.
- (332) Church G. M. The Personal Genome Project. *Molecular Systems Biology*. 2005.

- (333) Jonietz E. Personal genomes - Individual sequencing could be around the corner. *Technology Review*. 2001, 104: 30-30.
- (334) Church G. M. Genomes for all. *Scientific American*. 2006, 294: 46-54.
- (335) Berg P., Baltimore D., Brenner S., Roblin R. O., Singer M. F. Summary Statement of Asilomar Conference on Recombinant DNA-Molecules. *Proceedings of the National Academy of Sciences of the United States of America*. 1975, 72: 1981-1984.
- (336) Berg P., Singer M. The Recombinant-DNA Controversy - 20 Years Later. *Bio-Technology*. 1995, 13: 1132-1134.
- (337) Cook-Deegan R. M. The Alta summit, December 1984. *Genomics*. 1989, 5: 661-663.
- (338) Choudhuri S. The Path from Nuclein to Human Genome: A Brief History of DNA with a Note on Human Genome Sequencing and Its Impact on Future Research in Biology. *Bulletin of Science, Technology & Society*. 2003, 23: 360-367.
- (339) DeLisi C. Santa Fe 1986: Human genome baby-steps. *Nature*. 2008, 455: 876-877.

Biography

B.1. Personal

Name:

Faisal Reza.

Birth:

Dhaka Medical College and Hospital. Dhaka, Bangladesh. 1980.

B.2. Training

Undergraduate:

Massachusetts Institute of Technology (MIT). Cambridge, MA, USA.

S.B. Physics. 2003.

S.B. Science, Technology, and Society. 2003.

Minor. Biomedical Engineering. 2003.

Minor. Chemistry. 2003.

Thesis. Reza F. Human cloning: science, ethics, policy, society. 2003.

Graduate:

Duke University and Medical Center. Durham, NC, USA.

M.S. Biomedical Engineering. 2006.

Ph.D. Biomedical Engineering. 2010.

Certificate. Biomolecular and Tissue Engineering. 2010.

Certificate. Computational Biology and Bioinformatics. 2010.

Certificate. Biological and Biologically Inspired Materials. 2010.

Certificate. Computational Science and Engineering. 2010.

Dissertation. Computational molecular engineering nucleic acid binding proteins and enzymes. 2010.

B.3. Research

Reza F., Zuo P., Tian J. Protein interfacial pocket engineering via coupled computational filtering and biological focusing criterion. *Annals of Biomedical Engineering: Special Issue: Systems Biology, Bioinformatics, and Computational Biology*. 2007, 35: 1026-1036.

Reza F., Chandran K., Feltz M., Heinz A., Josephs E., O'Brien P., Van Dyke B., Chung H., Indurkha S., Lakhani N., Lee J., Lin S., Tang N.,

LaBean T., You L., Yuan F., Tian J. Engineering novel synthetic biological systems. *IET Synthetic Biology*. 2007, 1: 48-52.

Tan C., Reza F., You L. Noise-limited frequency signal transmission in gene circuits. *Biophysical Journal*. 2007, 93: 3753-3761.

Ma K-S., Reza F., Saaem I., Tian J. Versatile surface functionalization of cyclic olefin copolymer (COC) with sputtered SiO₂ thin film for potential BioMEMS applications. *Journal of Materials Chemistry*. 2009, 19: 7914-7920.

Reza F., Wang Q., Georgiev I., Donald B. R., Tian J. Automated and accurate engineering of a superior nucleic acid enzyme. *In revision*.

Reza F., Donald B. R., Tian J. Computationally engineered superior restriction endonuclease. *Invention disclosure submitted*.

Reza F., Georgiev I., Tian J., Donald B. R. Open-source computational redesign of nucleic acid binding proteins. *To be submitted*.

B.4. Teaching and service

Undergraduate Association Brass Rat Committee. MIT. Chair. 1999-2000.

Biomedical Engineering Society Chapter. MIT. Member-at-large. 2000-2002.

Undergraduate Research Journal (MURJ). MIT. Editorial Board. 2000-2002.

Bioinformatics, MIT. Teaching Assistant. 2003.

Biological and Bioprocess Engineering. MIT / MUST (Malaysia University of Science and Technology). Teaching Assistant. 2003.

Genomics and Computational Biology. Harvard University. Teaching Fellow. 2003.

Biological Sciences Graduate Student Symposium. Duke University. Organizer. 2005. Contributor. 2004, 2006-2009.

Genomic and Proteomic Technology. Duke University. Teaching Assistant. 2006.

Genetically Engineered Machines Program. Duke University. Co-organizer. 2006-2009.

Biomolecular and Tissue Engineering Chalk Talks. Duke University. Organizer. 2007-2008.

Introduction to Biomaterials. Duke University. Teaching Assistant. 2008.

Sigma Xi, The Scientific Research Society Chapter. Duke University. President. 2008-2010.

B.5. Awards and honors

Robert C. Byrd National Honors Scholar. US Department of Education. 1998-2002.

Institute Arts Scholar. MIT. 2000-2002.

Bioengineering Undergraduate Research Award. MIT. 2001.

Ronald H. Cordover Endowed Scholarship. MIT. 2001-2002.

Whitaker Research Award. MIT. 2002.

NIH-MIT Research Fellowship in Macromolecular Interactions. NIH. 2002.

Computational Biology and Bioinformatics Award. Duke University. 2004-2007.

NIGMS Biotechnology Predoctoral Fellowship. NIH. 2005-2007.

Program for Excellence in Science. American Association for the Advancement of Science. 2005.

Graduate Student Mini-Grant Award. Sigma Xi, The Scientific Research Society. 2007.

Kewaunee Best Poster Junior Graduate Student Award. Duke University. 2007.

Outstanding Graduate Teaching Assistant Award. Duke University. 2007-2008.

Conference Medal for Engineering. Sigma Xi, The Scientific Research Society. 2008.

Grand Challenges National Summit First Place for Health. National Academy of Engineering. 2009.

Kewaunee Student Service Award. Duke University. 2008-2009.

Annual Meeting Graduate Student Research Award. Biomedical Engineering Society. 2009.

Graduate School Conference Travel Fellowship. Duke University. 2009.

Conference Medal for Math and Computer Science. Sigma Xi, The Scientific Research Society. 2009.

Member. American Association for the Advancement of Science.

Member. Biomedical Engineering Society.

Member. Sigma Xi, The Scientific Research Society.

Member. Tau Beta Pi, The Engineering Honor Society.



*Sangram (Struggle),
circa 1976.*

– Zainul Abedin
Bangladeshi artist