

Appendix 7: More Support for Molecular Knowledge

Appendix 6 discussed the LacI gene experimental data, indicated that the analytical technique developed in chapter 4 assigns 0.42 bits to each amino acid, and then referred readers to the web site for more information.

This appendix will elaborate on the LacI experimental data and how it compares to SIFT and the techniques developed in chapter 4. The experimental data for LacI is complicated because very few organisms have this gene. Thus, to even apply the analytical technique, genes other than LacI must be included in the multiple sequence alignments. These genes encode proteins whose amino acid sequence is very similar to LacI, but their function may be slightly different. All of these genes belong to the LacI family.

Appendix 6 was able to apply the analytical technique with only 6 species because of the diversity of the organisms involved. Mouse ear cress is a plant, and *E. Cuniculi* is one the most primitive eukaryotes. More genes could have been included because many bacteria have the gene that encodes ADPG. The bacteria were not included in the appendix 6 because they did not alter the results (data not published). That is the knowledge calculated with these 6 species was the same knowledge that was calculated when 7 species of bacteria were included. *E. cuniculi* and Mouse ear cress ensure that the species were diverse and more diversity was not required.

Unfortunately, the LacI gene only exists in a few bacteria. So it is not possible to accumulate diversity with the same gene in many different organisms separated by billions of years of evolution. This means that for this alignment genes that are similar to LacI but have some other function must be included in the multiple sequence alignment. This is not the ideal situation. Nevertheless, because there is so little useful experimental evidence concerning the effects of random mutations are random sites, LacI is the next best candidate after AAG. The sequence alignment is found on the next page.

This alignment was created by finding the LacI in E. Coli at this web site: www.us.expasy.org/sprot/. At the bottom of the page for LacI in E.coli there is an option to run blast. This will search the database for similar sequences. Several sequences belonging to the LacI family were selected from the sequences that blast returned and clustal W was then used to create a multiple sequence alignment. Table A7 shows the results.

pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	bits
34	L	L	L	L	L	L	L	I	L	M	M	M	M	I	1.8
36	D	D	D	D	D	D	D	D	D	D	D	D	D	D	4
37	V	V	V	V	V	V	V	V	V	V	V	V	V	V	1.8
38	A	A	A	A	A	A	A	A	A	A	A	A	A	A	1.8
43	V	V	V	V	V	V	V	V	V	V	V	V	V	I	1.8
48	V	V	V	V	L	V	V	V	V	V	V	V	V	V	1.8
50	R	R	R	R	R	R	R	R	R	R	R	R	R	H	2.67
51	V	V	V	V	V	V	V	V	V	V	V	V	V	V	1.8
58	V	V	V	V	V	V	V	V	V	V	V	V	V	V	1.8
63	R	R	R	R	R	R	R	R	R	R	R	R	R	R	2.67
66	V	V	V	V	V	V	V	V	V	V	V	V	V	V	1.8
70	M	M	I	I	I	M	I	V	I	M	M	M	I	I	1.8
73	L	L	L	L	L	L	L	I	L	L	L	L	L	L	1.8
75	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	3.67
77	P	P	P	P	P	R	R	P	P	P	P	R	P	P	4
81	A	A	A	A	A	A	A	A	A	A	A	A	A	A	1.8
84	L	L	L	L	L	L	L	L	L	L	L	L	L	L	1.8
92	I	L	M	M	M	I	I	I	I	I	I	V	I	I	1.8
93	G	G	G	G	G	G	G	G	G	G	G	G	G	G	4
94	V	L	L	L	L	V	I	V	V	V	V	V	L	L	1.8
95	A	V	V	V	V	I	L	L	V	I	I	I	V	L	1.8
108	V	A	A	A	A	A	L	L	L	L	L	L	L	A	1.8
111	I	I	I	I	I	V	L	I	I	I	I	I	I	I	1.8
150	L	V	V	V	V	I	I	A	V	V	V	V	L	L	1.8
151	I	I	I	I	I	V	I	I	I	I	I	I	I	I	1.8
152	I	V	I	I	I	V	A	V	A	I	V	A	V	V	1.8
203	L	L	L	L	L	L	L	L	L	L	L	L	L	L	1.8
205	A	E	E	E	E	E	D	D	E	D	D	D	D	D	4
207	G	G	G	G	G	G	G	G	G	G	G	G	G	G	4
208	H	H	H	H	H	H	H	H	H	H	H	H	H	H	2.67
216	G	G	G	G	G	G	G	G	G	G	G	G	G	G	4
226	R	R	R	R	R	R	R	R	R	R	R	R	R	R	2.67

230	W	W	W	W	W	W	W	W	W	W	W	W	W	Y	3.67
234	L	L	L	L	L	L	L	L	L	L	L	L	L	L	1.8
251	W	W	W	W	W	W	W	W	W	W	W	W	W	F	3.67
272	A	A	A	A	A	A	A	A	A	A	A	A	A	A	1.8
273	M	I	V	V	V	I	V	I	L	V	V	V	V	I	1.8
275	V	V	V	V	V	V	V	A	V	A	A	A	A	A	1.8
278	D	D	D	D	D	D	D	D	D	D	D	D	D	D	4
281	A	A	A	A	A	A	A	A	A	A	A	A	A	A	1.8
282	L	L	L	L	L	L	L	L	A	L	L	L	L	I	1.8
283	G	G	G	G	G	G	G	G	G	G	G	G	G	G	4
284	A	V	V	V	V	V	V	L	V	L	L	L	V	V	1.8
285	M	L	L	L	L	L	L	I	L	L	L	L	I	L	1.8
287	A	A	A	A	A	A	A	A	A	A	A	A	A	A	1.8
300	S	S	S	S	S	S	S	S	S	S	S	S	S	S	2.67
301	V	V	V	V	V	V	V	V	V	V	V	V	V	V	1.8
303	G	G	G	G	G	G	G	G	G	G	G	G	G	G	4
304	Y	F	Y	Y	Y	Y	F	F	F	F	F	Y	F	F	3.67
305	D	D	D	D	D	D	D	D	D	D	D	D	D	D	4
306	D	D	D	D	D	D	D	D	D	D	D	D	D	D	4
315	P	P	P	P	P	P	P	P	P	P	P	P	P	P	4
317	L	L	L	L	L	L	L	L	L	L	L	L	L	L	1.8
318	T	T	T	T	T	T	T	T	T	T	T	T	T	T	2.67
319	T	T	T	T	T	T	T	T	T	T	T	T	T	T	2.67
320	I	I	V	V	V	V	V	V	V	V	V	V	V	V	1.8
328	G	G	G	G	G	G	G	A	G	G	G	G	G	G	4
357	V	I	V	V	V	I	V	V	V	V	V	V	V	V	1.8
359	R	R	R	R	R	R	R	R	R	R	R	R	R	R	2.67
361	T	S	S	S	S	S	S	S	S	S	S	S	T	S	2.67

Total number of bits = 152 (sum of column 16), LacI gene ~ 360 amino acids. So the molecular knowledge is $152/360 = 0.42$ bits per amino acid. This corresponds to on average 15 allowed amino acids per site.

Since experimental evidence suggests that 34% of all mutations deactivate LacI (appendix 6), the experimental data only allows 13.2 amino acids per site or 0.6 bits per site. Because this analysis included genes from the LacI family, it is not surprising that the analytical technique underestimates the molecular knowledge. The next page shows the results from SIFT.

Table A7: Allowed amino acids as predicted by SIFT

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1M	M																				1
2K	K																				1
3P	r	q	v	k	d	e	n	g	t	a	P	S									12
4V	c	e	k	p	f	s	m	a	T	I	I	V									12
5T	T																				1
6L	L																				1
7Y	c	w	m	p	i	g	t	q	n	r	v	s	k	d	a	I	f	H	Y	E	20
8D	D																				1
9V	V																				1
10A	A																				1
11E	I	p	g	n	t	s	a	q	d	k	R	E									12
12Y	c	w	m	p	d	e	g	n	q	i	t	s	v	a	k	I	f	R	H	Y	20
13A	A																				1
14G	G																				1
15V	V																				1
16S	S																				1
17Y	t	r	n	s	v	a	I	f	H	Y											10
18Q	Q																				1
19T	T																				1
20V	V																				1
21S	S																				1
22R	R																				1
23V	V																				1
24V	f	m	V	I	L																15
25N	N																				1
26Q	i	h	v	p	I	g	t	d	a	s	e	r	N	Q	K						15
27A	w	c	m	i	f	p	y	H	v	I	q	r	d	n	k	g	e	t	S	A	20
28S	r	I	q	n	d	k	v	e	t	g	S	P	A								13
29H	m	f	i	v	y	p	I	t	a	q	g	s	e	d	r	H	N	K			18
30V	V																				1
31S	S																				1
32A	m	h	i	p	v	I	g	n	r	t	q	d	S	A	K	E					16
33K	m	y	i	h	v	p	I	g	d	n	t	s	q	e	A	R	K				17
34T	T																				1
35R	R																				1
36E	y	m	i	h	v	p	I	g	t	q	s	a	d	N	k	R	E				17
37K	q	R	K																		3
38V	V																				1
39E	E																				1
40A	f	c	m	y	h	i	I	n	p	v	r	t	d	Q	g	k	s	e	A		19
41A	t	g	S	A																	4
42M	a	f	v	I	M	I															6
43A	h	i	I	r	v	n	p	t	g	q	s	k	d	A	E						15
44E	c	y	m	f	h	p	i	g	n	r	v	L	t	q	s	d	k	A	E		19
45L	L																				1
46N	c	m	f	i	v	y	I	p	h	t	q	a	e	R	s	G	k	d	N		19
47Y	Y																				1
48I	h	g	c	d	n	p	m	y	q	f	e	R	s	k	T	a	I	I	V		18
49P	f	m	y	i	h	v	d	I	n	t	g	s	e	q	a	k	R	P			18

50N	N	1
51R	h d p v g l n e a q s k T R	14
52V	e c k s p f m t i A L V	12
53A	A	1
54Q	a e k R Q	4
55Q	l v p r g n d k t e a S Q	13
56L	L	1
57A	k e c m g p s t l i A V	12
58G	q r v e p k d n s a T G	12
59K	q R K	3
60Q	f m y i h v p g l d n a s e T k R Q	18
61S	h c m y f q r d p e n k g l l v a t S	19
62L	w c p d n g q e M k r s T H i a v f y L	20
63L	h c y d g q r e p n f m k a i s v L T	19
64I	a f v M I L	7
65G	G	1
66V	t f a m i V L	7
67A	f m t i A L V	7
68T	v s A T	4
69S	n a S T	4
70S	i h v l p r g t q n a k D S E	15
71L	L	1
72A	p t g S A	5
73L	s t e r a k f m v Q i L	12
74H	l f Y H	5
75A	s G A	4
76P	P	1
77S	t g A S	4
78Q	h p t a r g s e k d N Q	12
79I	m a I T I V	6
80V	n g y r q c e k p f s m t i A L V	17
81A	y n d c q r e k p f m g t s i v L A	18
82A	s G A	3
83I	I V I	3
84K	s a q d r E K	7
85S	f m y i h v p l g d n t a Q S e R k	18
86R	w c m p d n g i q e R h v T s k A f I Y	20
87A	A	1
88D	w c m f i y p v l H q g t R a n k S D e	20
89Q	c m f y h l p v g l r n t Q s a k D e	19
90L	c m y h f p i g n r q t v L s d k A E	19
91G	G	1
92A	c w d p m e q n k g r s t h l A v f I Y	20
93S	w c m p i h F v g r Q I d y N t k a e S	20

94V	I I V	3
95V	h c d n q g r p e y k m f t S a i L V	19
96V	I M V I	4
97S	t g A S	4
98M	n c d q r e k f p t g s i v l A M	17
99V	L V I	3
100E	m h i v p l g n t q R a S k D E	16
101R	h i v p l g n t q d a k R S E	15
102S	h m f y c r q i d n l P e k g V t S a	19
103G	w c m f h y i l p r V q t k n G D e s a	20
104V	c h m f y p r q N d g k e l i s t A V	19
105E	h v l p g r t s a k d N Q E	14
106A	k n g A T S	6
107C	e k p y s m f t a C l v l	13
108K	f y m h i p v l g t N d r s Q A e K	18
109A	c m i f v y l p H r q t A k g s e D N	19
110A	t g S A	4
111V	m t I A I V	6
112H	c m i f v p y l t a q e g s H d k R N	19
113N	y r p q t a k e s H g N D	13
114L	v m i F L	6
115L	d h g p n y s e t q a r f M K v i L	18
116A	h i l v p r q n g k t e D A S	15
117Q	Q	1
118R	W c m i p d v f h t e G n l q s y a R k	20
119V	V	1
120S	f i c y l v h r p q t k e a S G D n	18
121G	y v h l p t q e d s n r a K G	15
122L	L I V	3
123I	I V I	3
124I	f m t I A V I	7
125N	w c h p y m f q r e g t v a d k s i N L	20
126Y	m s t a f l i Y V	9
127P	P	1
128L	L	1
129D	m y i v l p r H g t q n a k S D E	17
130D	f c y m h i l p v r g q n k e D A s T	29
131Q	c f y m h i v l p r G t n Q s D A k e	19
132D	c f y m h i p l V g n r q T s D a k E	19
133A	n y g r q c e k p f s m t i A L V	17
134I	c f y m h l p l v g n r t Q s d A k E	19
135A	c m f i y p H v l g d n t s q A e r K	19
136V	L V I	3

137E	f y c m h i l v p r n g q t d k S A E	19
138A	f m c y h i l n p v r t d g k s Q e A	19
139A	c m f y h i p L g v n r q D T s A k e	19
140C	m f C i y l v r h q p e k t a g s d N	19
141T	c m f y h i p L g v n r q D T s A k e	19
142N	e s g N D	5
143V	l a i T V	5
144P	q i r l d e k n g v a s T P	14
145A	r q f e k g m p s C l t i A V	15
146L	m i V L	4
147F	w t m a y i l V F	9
148L	f i M V L	5
149D	r h p q t k a e s n G D	12
150V	h y f m c r d n q e k p l i g t S V A	19
151S	h r p q k a t e g n S D	12
152D	f i c y l v h r q t k P a e s G n D	18
153Q	c w p m d e n g Q k r i t s v A H l f Y	20
154T	f m y h c i l r q v p e k d n G a S T	19
155P	c f y m h i v l P g R q n t D a S k e	19
156I	l F V l	4
157N	w c p m d q g e N r k i t h S v a F L y	20
158S	l y r p h q t k a e d S G N	14
159I	t l A l V	5
160I	w c h d p g n q r e y k M f S a t l v l	20
161F	w t m a y i l V F	9
162S	p r h q a t k e g S N D	12
163H	w c m i f v y l g H t P n s a d Q r e k	20
164E	m h i v l p g r n t q s A D K E	16
165D	w c M i p v f r H l q t g y k a s n e D	20
166G	G	1
167T	e n k l i p g v s A T	11
168R	R	1
169L	y r q e k p g s t f m i v A L	15
170G	d e k v t p G S A	9
171V	l i V	3
172E	h i v l p g n t r s q d A K E	5
173H	s v a l f Y H	7
174L	L	1
175V	m t a f Y l i V	8
176A	v r n p t g s q k d A E	12
177L	h p g y n s e t q a f m v i k R L	17
178G	G	1
179H	H	1

180Q	d v g p h l s n t a e Q K R	14
181Q	f m y i v p l g n t d s a H Q k R E	18
182I	l	1
183A	s G A	3
184L	d p g n e s q t r k y a H m f v i L	18
185L	m v l L	4
186A	d p v e n k g T S A	10
187G	G	1
188P	P	1
189L	M f y i p h d g v n t s a L e q R K	18
190S	c f y m h i v p l g r n Q D a T k S e	19
191S	m c W h i q r e p v k l d f y n g a t S	20
192V	l l V	3
193S	h r p q k a e g t n D S	12
194A	A	1
195R	c m f y i h v p l g t q d s N e A R k	19
196L	h y g p n s r t q f m d k a v i E L	18
197R	R	1
198L	c d w p e n q k m g r t s h i v f A Y L	20
199A	m h i v l p g n t r s q d A K E	16
200G	p t d n a S G	7
201W	W	1
202H	c w d m p g n i s t e v H f a q y L R k	20
203K	h i v l p g n t r s q d A K E	15
204Y	w c m p i g h n r q f d v l k T s Y A E	20
205L	L	1
206T	m h i l p g r n v q d k s T E A	16
207R	c y f m h i p v l g n t R q D s A K e	19
208N	w c m p i v h f g l t r Q N Y s a d k E	20
209Q	f m i y h v p l t d n s a G e Q R k	18
210I	m v l L	4
211Q	c f m y h i l v r P g n d Q k e t a S	19
212P	t v g s A P	6
213I	D y k e s m f t a l l V	12
214A	h m c f y i r n q d p k e L v g t S A	19
215E	h c g n y d p r q s f m k E t a L i V	19
216R	y R t m f a L V l	9
217E	y v l p t g r s n a q k d H E	15
218G	G	1
219D	k e s g N D	6
220W	W	1
221S	y l v h r p q k a t e g n D S	15
222A	A	1

223M	w c M f h i y p v l n r G Q d t s k e A	20
224S	S	1
225G	G	1
226F	I F Y	3
227Q	w c m i f p H v y l r Q t G a n d k S e	20
228Q	w c m f h p l y v G n l d Q t r s e a k	20
229T	h f m y i c l r q n v d k e p T G s A	19
230M	t a Q f v M i L	8
231Q	y m i h v p l g n t d s r e A Q K	17
232M	f v i M L	5
233L	g s t f m v i A L	9
234N	w c m f i h y p V l G N t q d R s a e k	20
235E	d Q E	3
236G	c f m y h i p v l G R q n d a e T k S	19
237I	w c P m h g d N q r f y s e t k l a v L	20
238V	c m f y h i p g V d l n t s a Q e R k	19
239P	w c d P m n q e g r k h s t F l A y v l	20
240T	n a T S	4
241A	A	1
242M	I M I V	4
243L	a m i F V L	6
244V	V	1
245A	s G A	3
246N	N	1
247D	D	1
248Q	Q	1
249M	M	1
250A	A	1
251L	L	1
252G	G	1
253A	s t l i A V	6
254M	i M L	3
255R	y i h v l p d g n q e t a k S R	16
256A	A	1
257I	a y m v l F L	7
258T	m c f i p l q y r v g H e d k n a S T	19
259E	a k d Q E	5
260S	w c m p i q r v h f e k l g d t N A Y S	20
261G	f c i y h v l p t r e k s n d a Q G	18
262L	m i V L	4
263R	c f m y h i d n p l v t q e g s k A R	19
264V	V	1
265G	i y h c l q r v k e t n d s G a P	17
266A	c f y m h i p v l n G t q R d s A k E	19

267D	g t n p s a k Q D E	10
268I	c g d n p q y r s e K M f t a l l V	18
269S	S	1
270V	V	1
271V	l V l	3
272G	G	1
273Y	w l F Y	4
274D	D	1
275D	D	1
276T	i l p g v d r n e a k s Q T	14
277E	w c m i P g h v f n t l r s Q Y a k d E	20
278D	a k q D E	5
279S	t g A S	4
280S	c r q l n d e k v t p G S A	14
281C	d p k e q m n C g r t s h w i a v l F Y	20
282Y	w l Y F	4
283I	c w d p m e q g n k r h s a T l f l v Y	20
284P	P	1
285P	k e t v s g A P	8
286L	L	1
287T	T	1
288T	T	1
289I	l l V	3
290K	m y i h v p l g d n t a q e S K R	17
291Q	h y p g n d s t Q e f a m r v k i L	18
292D	D	1
293F	m v y i L F	6
294R	i h v l p g t n R s q a k D E	15
295L	h y g p n s r t q f m d k a v i E L	18
296L	d g y n p s t e r a k Q f m v i L	17
297G	G	1
298Q	t a e Q R K	6
299T	y m h i v p l g n q d s a T R k E	17
300S	r q n l d e k v t p C G S A	14
301V	V	1
302D	c f m y h i v l p g q R n a T s D k e	19
303R	d y p f h g m n s t e v a i q L k R	18
304L	f v i M L	5
305L	L	1
306Q	c f y m h i p v l g n t d Q R A e S k	19
307L	h c y d n f m r q p g e k s i T L A v	19
308S	w c h y q d r p m n e f k g i v t S a L	20
309Q	m h i v l p g n t R s a Q k D E	16

310G y r p h q t a k e s G D N 13
 311Q m h i v l P G r n t Q s a d k E 16
 312A l r q d p v e k n g T S A 13
 313V c h p g f y n d r s M Q k e t i l A V 19
 314K m y i h v p d l g n t s A q e R K 17
 315G h c i l r v q p e k d n a T G S 16
 316N p g e k d a N T S 9
 317Q c m f y h i p v g L r n d Q a k e T S 19
 318L a f m L V l 6
 319L m i V L 4
 320P r t n g s q k a d E P 11
 321V h y f c m n r d q g k e P l s i T A V 19
 322S m h i v p l g n t d q a S R E K 16
 323L m v l l 4
 324V l l V 3
 325K f d g n m p q s e r t l a K i V 16
 326R R 1
 327K c m f i y p H v l g t q r d s N A e K 19
 328T a T S 3
 329T T 1
 330L h m c f i y n q r d e v k t p L s G A 19
 331A w c m f h y P l g n d l q v R t s e k A 20
 332P w c m i f v y H l P g t n s a d Q r e k 20
 333N c f m y h i l v p r q g d k N e t A S 19
 334T f c y m h i l p v g q n r d a e K S T 19
 335Q h i v l p g n r s a d k T Q E 15
 336T w c m i f v p l y r q H a k g T e s n D 20
 337A f c m y h i d n l p v q t e g s R k A 19
 338S y l v h r p q k a t e g n D S 15
 339P d n y q P e r g k s t f m a v i L 17
 340R d v p h g l s n t a e k Q R 14
 341A m h i v p l g n t r q s d A K E 16
 342L f m v l l 5
 343A i l p k e g v s T A 10
 344D v l p g r n t s q a K D E 13
 345S f y m i h v l p r g q t a N k d S E 8
 346L L 1
 347M d p g h n s y e t q a M f R v k i L 8
 348Q c m f y h i p v g L r n t Q s a D k e 19
 349L f m v l l 5
 350A i l p k e g v s T A 10
 351R d h p g l n s t a e k Q R 13
 352Q i v d h p g l n s t a e k Q R 15

353V	a f m i V L	6
354S	c f y m h i l p v n r g d t k e Q S A	19
355R	R	1
356L	L	1
357E	E	1
358S	S	1
359G	G	1
360Q	Q	1

The first column lists the site position and the amino acid for the protein encoded by E. Coli LacI. The other letters represent allowed amino acids. The far right column lists the number of amino acids allowed at each site. This average is roughly 10.4 amino acids per site or .94 bits of knowledge per amino acid. This is more than double the predicted knowledge using the techniques developed in chapter 4. Thus, the techniques used by this book always seem to assign less molecular knowledge to a protein than SIFT.

The main differences between these two analytical techniques are in the amino acid groupings and in the rules that predict which amino acids that should be allowed. Both allow all amino acids that natural selection allows. One of the main differences is that SIFT only allows one amino acid when a site is absolutely conserved by natural selection. The techniques of chapter 4 allow all amino acids from the same group in this case. The other primary difference is that the chapter four technique assigns zero bits to most of the protein (table A7 is much shorter than the SIFT results because positions with zero bits are not shown). Taken together these two difference explain why the techniques used by this book will always assign less molecular knowledge to a protein than SIFT.

<http://www.theory-of-evolution.net>