Appendix 1: Shannon Entropy and Information

In chapter 1, Shannon entropy is mentioned in passing. Shannon entropy is intimately tied to information theory. For readers who may read more about information theory elsewhere, this section will prove useful.

Information is transmitted in symbols. These symbols are most frequently letters. The English alphabet is a group of symbols. These symbols can be arranged to contain information. The average information per symbol is called Shannon entropy. It is almost always represented in the literature by a capital H.

H = Shannon Entropy

If all symbols are equally probable, the Shannon entropy is defined as:

H = 3.32 x log(number of possible symbols) Eq. 1

or for readers with a base 2 log function on their calculator

$H = \log_2$ (number of possible symbols)

For example, if 26 blocks are labeled with letters from the alphabet and then drawn from a hat, the average information per symbol is H = 3.32 x log (26) = 4.7 bits per letter.

Now suppose that 102 blocks with the letter *Z* are added to the hat in the previous example. The hat now contains 128 blocks, but most are labeled with a *Z*. The Shannon entropy is now the weighted average of the contents in the hat. The odds of pulling the letter A from the hat are 1 in 128. So from equation 2, in chapter 1, the information associated with drawing an A is 3.32 x log (128/1) = 7 bits. With the exception of the letter Z, all other letters have the same odds. Thus, observing any letter, A-Y, results in 7 bits of information.

The letter Z will be drawn from the hat 102 times with every 128 tries. This corresponds to 1 time in 1.25 tries, and the corresponding information is 3.32 x log (128/102) = 1.6 bits.

Shannon entropy is the weighted average. 26 symbols contribute 7 bits and 102 contribute 1.6 bits. So H is calculated as follows: (102/128) x 1.6 + 26/128 x (7) = 1.275 + 1.42 = 2.7 bits per symbol. This means that a code exists that can transmit the contents of the hat using on average only 2.7 bits per symbol.

> *Question:* How much information is carried by this message: “AAAAAAAA” if it is drawn from the hat with 128 letters?
> *Answer:* Each A contributes 7 bits, so this message contains 8 letters x 7 bits per letter = 56 bits of information.
> *Question:* what are the odds of drawing it from the hat?
> *Answer:* 1 in 2^{56} or 1 in 7.2 x 10^{16}.
> *Question:* on average how much information will most 80 letter messages drawn from this hat contain?
> *Answer:* 80 letters x 2.7 bits per letter = 216 bits.
> *Question:* How much information is carried by this message: “ZZZZZZZZ”?
> *Answer:* Each Z contributes 1.6 bits. So this message contains 8 letters x 1.6 bits per letter = 12.8 bits of information.

Shannon entropy only works for very long messages. Short messages may or may not be accurately represented by Shannon entropy. The information content of a short message may be much higher or much lower. The examples above illustrate this concept. Long messages will always converge to the average information per bit. This is why Shannon entropy is useful.

Appendix 2: Relative Entropy and Information

This section is best read immediately after chapter 4. The techniques used in chapter 4 to calculate information are different from those used by most authors. Most use relative entropy.

Suppose that a sequence alignment for a protein gives the followings results (only the first two amino acids in the sequence are shown):

Chicken	AlaVal..
Man	AlaVal..
Dog	AlaVal..
Lizzard	AlaAla..
Fish	AlaAla..
JellyFish	AlaAla..

The information in the first position is easy to calculate. Four out of 64 possible codons specify alanine. So the information is by equation 2 in chapter 1 as follows:

Information = 3.32 x log (64/4) = 4 bits.

The information of the second position is also easy to calculate. Four codons specify alanine and 4 specify valine. The information is as follows:

Information = 3.32 x log (64/ 8) = 3 bits.

These calculations are in full agreement with the techniques described in chapter 4.

Now consider a different alignment.

Chicken	AlaVal...
Man	AlaVal...
Dog	AlaVal...
Lizzard	AlaVal...
Fish	AlaVal...
JellyFish	AlaVal...
Oak Tree	AlaVal...
E coli	AlaAla...

The information content of the first position has not changed. It is still 4 bits, but what about the second position? The same two amino acids are present, but valine is found in 7 of the 8 sequences. Intuitively, it would seem that the second position in this alignment contains more information, and it does. To calculate the information for the second position, the formula for relative entropy must be used. The formula for relative entropy when two amino acids appear in the same alignment column is as follows:

$$\begin{array}{l}\text{Relative}\\\text{Entropy}\end{array} = \begin{array}{l}\text{Frequency}\\\text{amino acid 1}\end{array} \text{ x } 3.32\text{x log} \left[\frac{\begin{array}{l}\text{Frequency}\\\text{amino acid 1}\end{array}}{\begin{array}{l}\text{Expected Frequency}\\\text{of amino acid 1}\end{array}} \right]$$

$$+ \begin{array}{l}\text{Frequency}\\\text{amino acid 2}\end{array} \text{ x } 3.32\text{x log} \left[\frac{\begin{array}{l}\text{Frequency}\\\text{amino acid 2}\end{array}}{\begin{array}{l}\text{Expected Frequency}\\\text{of amino acid 2}\end{array}} \right]$$

In this case, the relative entropy is as follows:

Relative Entropy =
1/8 x 3.32 x log [(1/8)/(4/64)] + 7/8 x 3.32 x log [(7/8)/(4/64)]

Relative Entropy = .125 + 3.33 = 3.46 bits.

Relative entropy is a measure of information. In fact, the actual information at position 2 is the relative entropy.

Relative Entropy = Actual Information = 3.46 bits.

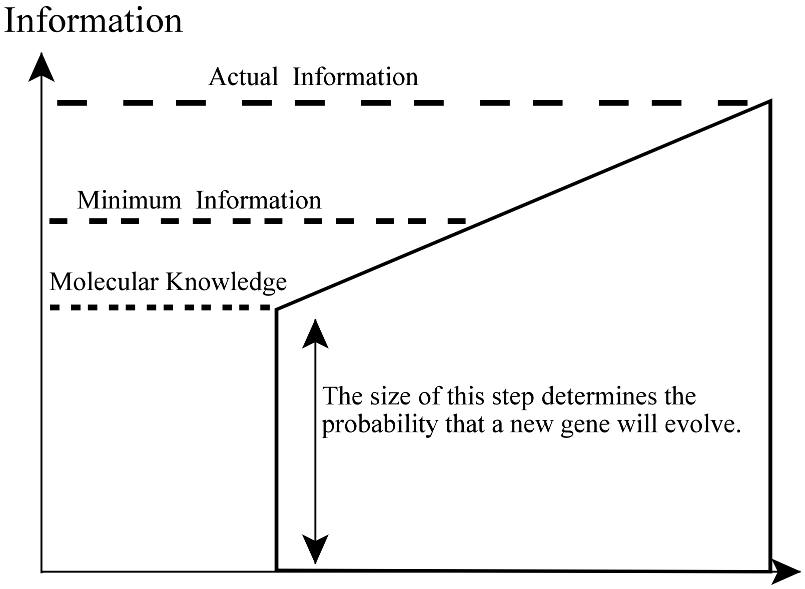
Now consider what happens when the equation for relative entropy is applied to the second position in the first set of sequences. In this set of sequences, valine occurs ½ of the time, and alanine occurs ½ of the time.

Relative entropy =
½ x 3.32 x log [(½)/(4/64)] + ½ x 3.32 x log[(½) /(4/64)] =
1.5 + 1.5 = 3 bits.

So relative entropy gives the same results as the technique used in chapters 4 and 5 when the amino acids in a column all occur at the expected frequency. The expected frequency is set by the underlying probabilities associated with each amino acid arising by chance.

Why not use relative entropy? Relative entropy is always greater than or equal to the information calculated with the techniques used in this book. These techniques always calculate the minimum possible information in the gene or protein today. The true information which can only be found by using relative entropy will always be higher. Relative entropy depends on the distribution of amino acids, and no such distribution can exist for the very first proteins. So while relative entropy is the correct method to measure information, it is not the best method to model the evolution of new proteins and genes.

Fortunately, with the simple assumption that all amino acids in a distribution occur at the expected frequency, the equation for relative entropy simplifies to those introduced in chapter 1 (pages 24-25). These two equations always calculate the minimum possible information in any distribution. Thus, the technique used in chapters 4 and 5 calculates the minimum possible information by assuming that all allowed amino acids in any given column are found at the expected frequency. Under some circumstances, this value can be related to a probability for evolution (chapter 5). But most of the time, this should never be done. Molecular knowledge defines the minimum information required for a selective advantage to be realized. It can almost always be related to a probability for evolution. The size of this first vertical step determines whether or not chance will create the required knowledge. The graph below shows how these quantities are related.


Information
Actual Information
Minimum Information
Molecular Knowledge
The size of this step determines the probability that a new gene will evolve.

Appendix 3: Math Review

Base 10 is composed of 10 integers. These are 0,1,2,3,4,5,6,7,8, and 9. Using these 10 integers, any number can be expressed quite easily. Base 10 is called base 10 because it uses 10 unique digits to express numbers. Most math problems use base 10.

Base 2 is composed of 2 integers. These are 1 and 0. Just like with base 10, any number can be expressed in base 2. Computers use base 2 for mathematical calculations. The following table shows how to count from 0 to 7 in both systems.

Base 10	Base 2
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

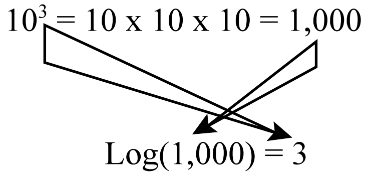
Base 2 is very useful for information theory because each digit in base 2 contains exactly 1 bit of information. So when information was defined, engineers naturally decided to define it in terms of base 2.

Exponents

Exponents indicate that a number should be multiplied by itself a certain number of times. In the following equation: $10^3 = 10 \times 10 \times 10 = 1000$, the exponent is 3 and in this equation $2^4 = 2 \times 2 \times 2 \times 2= 16$, the exponent is 4. Any number multiplied by itself zero times is by definition equal to 1, so $10^0= 2^0 = 10,000^0 = 1$.

Logarithms

The symbol log on a calculator indicates a logarithm. Logarithms are the inverse operation of exponents. For example, if the exponent 3 is used to raise 10 to a power the result is 1,000. If the log function is used to find the logarithm on 1,000, the result is equal to the exponent which in this case is 3.

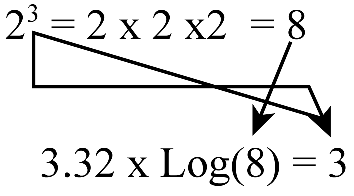

$10^3 = 10 \text{ x } 10 \text{ x } 10 = 1{,}000$
$\text{Log}(1{,}000) = 3$

Logarithms must assume a base. The log function on a calculator assumes that the user wants to calculate the logarithm for base 10. The equation below does not work because the log function is only the inverse when the number 10 is raised to a power by the exponent.

$$2^3 = 2 \text{ x } 2 \text{ x}2 \ = 8$$

$$\text{Log}(8) = 0.9 \neq 3$$

Because information theory uses base 2, the exponents always operate on the number *2*. The log function can be used to find the inverse of this operation if the answer is multiplied by a conversion factor, 3.32. This only works for exponents raising the number *2* to a power.


$2^3 = 2 \text{ x } 2 \text{ x}2 \ = 8$
$3.32 \text{ x } \text{Log}(8) = 3$

Addition with logarithms is the same operation as multiplication. The following equation will multiply 100 x 100 using logarithms.

Log (100) + Log (100) = 2 + 2 = 4

The number *4* is now used to raise the number *10* to a power:

10^4 = 10 x 10 x 10 x 10 = 10,000

Notice that 100 x 100 also equals 10,000. Today with handheld calculators logarithms are seldom used to multiply large numbers, but the property of logarithms that turns multiplication into addition is still very useful for information theory. It is this property of logarithms that led scientists to define information using logarithms because it allows information to be added without having to worry about multiplication.

Consider the following two sentences:

I like the dog.

I like the cat.

Both sentences use 15 characters, 11 letters, 3 spaces, and 1 period. If the sentences are constructed by drawing blocks labeled with the 26 letters in the alphabet, 1 space and 1 period, the odds of drawing either sentence is 1 in 5.1 x 10^{21}. The odds of drawing the first sentence followed by the second are 1 in (5.1 x10^{21}) x (5.1 x 10^{21}) = 1 in 2.6 x 10^{43}. Intuitively, these two sentences together should have twice as much information as either one by itself, and this is where the log function comes into play.

3.32 x Log (5.1 x 10^{21}) = 72 bits for either sentence alone.

3.32 x Log (2.6 x 10^{43}) = 144 bits for both sentences

Appendix 4: A Review of Yockey's Approach

Today, many scientists apply information theory to molecular biology, but only a few have tried to use information theory to answer the most important question. Is evolution possible? Yockey was probably the first. Because so few scientists are trying to answer this question, a consensus as to how to assign information has not been reached.

Yockey assigns information to proteins by a technique that is very different from the one developed in chapter 4. His technique relies on both Shannon entropy and conditional entropy. He models the transfer of information from DNA to proteins and assigns information to this transfer by treating mutations as a source of noise. Unfortunately, his analysis fails to consider natural selection, and so the information that he calculates is incorrect. Natural selection weeds out harmful mutations. In a communication system, this is analogous to a person receiving an unintelligible message and then asking for the message to be re-sent. The equations that Shannon developed to model noise in communication systems do not apply if a person on the receiving end inspects the incoming messages for errors and discards any messages that contain errors.

This book did not use Yockey's technique because the information assigned by his technique cannot be related to a probability for protein evolution. For a full development of his technique, refer to the two references at the end of this appendix.

Yockey's technique divides every sites in a protein into two categories, absolutely conserved and not absolutely conserved. An absolutely conserved amino acid is one that never changes. For example, if column 20 in a multiple sequence alignment is always a glycine, then position 20 in the alignment is absolutely conserved. If more than one amino acid is found in column 20, then it is not absolutely conserved. Yockey's technique sets the information content of any absolutely conserved amino acid equal to the Shannon entropy, 4.14 bits.

So using this technique, a peptide composed of 10 methionines has the same chance of evolving as one composed of 10 serines. This is the drawback of using Shannon entropy. Methionine is only specified by 1 codon, and serine is specified by six. Assuming random mutations, serine should arise by chance six times as often as methionine.

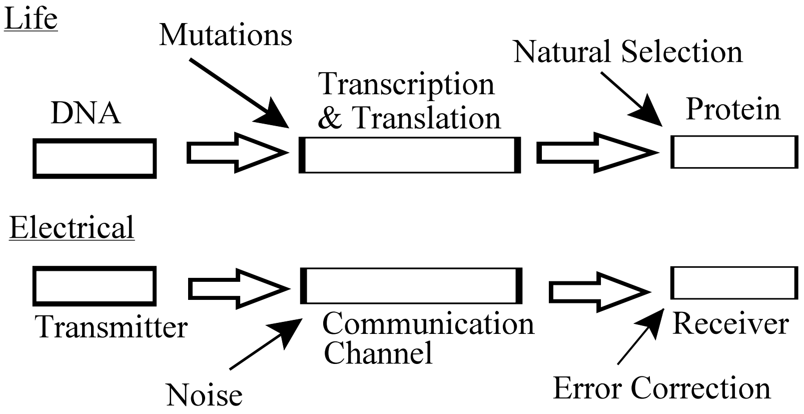
For reference, the Shannon entropy, assuming the genetic code, is calculated below. To follow Yockey's method exactly only the 61 codons that do not terminate the peptide chain are allowed.

amino acid group	expected frequency	information	total (bits)
6 codons (3 amino acids)	29.5%	3.34	0.99
4 codons (5 amino acids)	32.8%	3.93	1.29
3 codons (1 amino acid)	4.9%	4.34	0.213
2 codons (9 amino acids)	29.5%	4.93	1.45
1 codon (2 amino acids)	3.3%	5.93	0.195
			4.14

Example calculation: Three amino acids are specified by 6 codons. The information acquired when one of these is observed is as follows: information = 3.32 x log (61/6) = 3.34 bits. Because these amino acids should arise by chance 6 times in every 61 tries and because there are 3 of these amino acids, the expected frequency is (3 x 6)/61 = .295 or 29.5%. The last column is the product of the expected frequency and the information for each amino acid group (each row). The sum of all entries in the last column is the Shannon entropy.

Because proteins are short messages, Shannon entropy is almost never a true measure of a specific protein’s information. A typical protein may only have 30 absolutely conserved amino acids, and the probability for these amino acids arising by chance might be very different from the number calculated using Shannon entropy.

Yockey’s calculations for amino acids that are not absolutely conserved run into another issue. He models mutations as noise, and then applies the techniques developed by Shannon to calculate information transfer through a noisy communication channel. The figure below illustrates the similarity between information transfer in life and in electrical communication systems.



In electrical systems, conditional entropy models the amount of information lost due to noise. For example, suppose the transmitter transmits the results of a trapped scientist experiment. The two results are heads or tails. If the channel is noisy, when heads is transmitted tails might be received. This error reduces the rate of information transfer. Conditional entropy models how much information is lost. Mutations in life have the same effect as noise in communication systems, so mutations can be modeled as a noise source.

The problem with this approach is that natural selection does not allow harmful mutations to survive. So if a mutation creates a non-functional protein, the mutation will be removed from the population by natural selection. In communication systems, error correction is responsible for this function. With error correction, conditional entropy can no longer be used to accurately model the information lost due to noise. Likewise, because of natural selection, conditional entropy does not model the effect of mutations on information transfer.

Natural selection cannot ensure that only allowed messages are transmitted, but it does ensure that only allowed messages survive. And the neutral theory of evolution (see chapter 4) predicts that many if not all of the allowed amino acid substitutions in the final protein will be observed if the same proteins in many diverse species are analyzed. Thus, the information content of the final protein does not depend on the information transferred from DNA, and conditional entropy is not needed for this analysis. The equations to calculate information in chapter 1 are applicable, and they may be applied using the techniques introduced in chapters 4 and 5.

Furthermore, the techniques used in this book always maintain a strict one to one relationship between probability space and information space. In other words, probability theory and information theory both yield the same results. Yockey's approach does not preserve this one to one mapping. So the odds that he calculates for protein evolution are different than the odds that one would calculate using probability theory.

References:

1) Yockey, Information Theory, Evolution and the Origin of Life, 2005.

2) Yockey, Information Theory and Molecular Biology, 1992.

3) Shannon, A Mathematical Theory of Communications, 1948.

Appendix 5: A Review of Schneider's Approach

Tom Schneider has a very good web site of information theory. He has proposed a different way to assign information to a protein. His suggestion is to use relative entropy with the assumption that all amino acids are equally probably in the final protein. The resulting equation is shown below.

Let F1 = frequency of amino acid 1, F2 = frequency of amino acid 2 and so on. The last amino acid found at a given position is thus Fn.

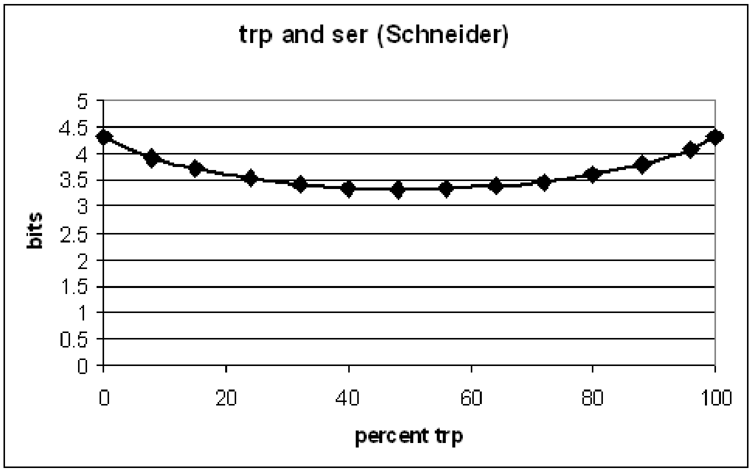
$$\text{Information} = \log_2 (20) + F1 \text{ x } \log_2 (F1) + F2 \text{ x } \log_2 (F2) \\ .....+ Fn \text{ x } \log_2 (Fn)$$

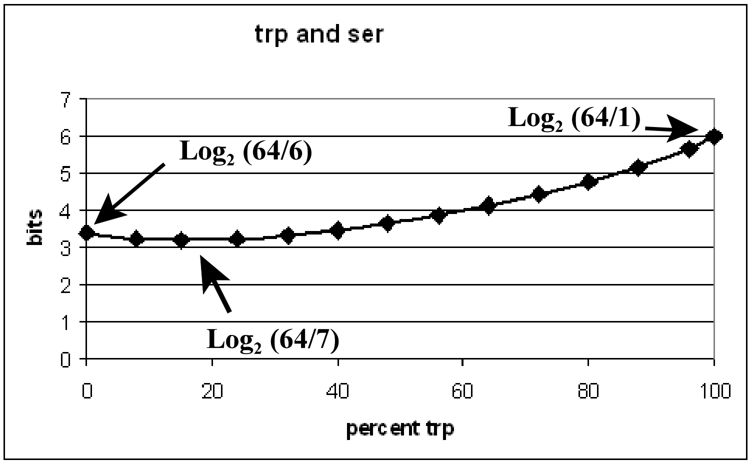
The first term is the maximum possible average information for each site within the protein. The terms that follow all are negative, and so these reduce the amount of information at each site (except when they are zero).

Using this equation, any column in a multiple sequence alignment that contains 1 and only one amino acid is assigned 4.32 bits. This assignment does not depend on the amino acid. So a column with only serine (6 codons) is assigned 4.32 bits, as is a position with only methionine (1 codon). Clearly, there is a problem with this approach. Serine is 6 times as likely to arise by chance as methionine. Assigning both of these columns the same information destroys any hope of relating the information in a protein to a probability of evolution. This problem is very similar to what was observed in appendix 4 with Yockey's technique.

Information theory works best when the messages are very long, and proteins are not long messages. Therefore, the average information per column (in the multiple sequence alignment) for all proteins in life is not necessarily equal to the information found in a specific column for any given protein.

The graphs below show how the information for a conserved site (that allows two amino acids) changes with the frequency of the allowed amino acids. The second graph uses relative entropy. The first graph uses Schneider's equation. Notice that the Schneider's equation never calculates information in such a way that it can be related to a probability of evolution. In contrast, relative entropy gives exactly the correct solution at the three distinct points shown in the graph.





The minimum observed information in the second graph is precisely the information predicted by using equations 1 and 2 introduced in chapter one. That is if trp and ser are found in a specific column in the multiple sequence alignment, then the column contains 3.32 x log (64/7) bits of information which of course is equal to $\log_2$ (64/7). It is a trivial exercise to show that the two endpoints in the second graph are also correct. Thus, relative entropy is the best measure of protein information. It is the only available approach that preserves the relationship between information and probability. It is preferable to both Schneider's and Yockey's techniques as these other two techniques can assign more information to a protein that is more likely to evolve. This is contrary to the definition of information as originally proposed by Shannon.

References:

1) Hertz, Stormo, Identifying DNA and protein Patterns with Statistically Significant Alignments of a Multiple Sequences, Bioinformatics, 1999.
2) http://www-lmmb.ncifcrf.gov/~toms/
3) Schneider, Stormo, Gold, Ehrenfeucht, "Information Content of binding Sites on Nucleotides Sequences," J. Mole Biology, 1986.
4) Schneider, Stormo, "Excess Information at Bacteriophage T7 Genomic Promoters Detected by Random Cloning Techniques," Nucl. Acids Res, 1989.
5) Schneider, "Measuring Molecular Information," Journal of Theoretical Biology, 1999.
6) Shannon, Weaver, The Mathematical Theory of Communication, 1964.

Appendix 6: Experimental Support for Molecular Knowledge

To calculate the molecular knowledge in proteins, the previous chapters have relied on natural selection's ability to preserve critical amino acids. Because natural selection will select against functional proteins that are slightly deleterious, not all functional proteins are expected in the online databases. To compensate for this effect, chapter 4 divided amino acids into groups based on their chemical properties and size. This chapter also proposed a model that allows amino acids from the same group at conserved sites within a protein even if these amino acids are not found in the databases. This model also assumes that any site that contains amino acids from more than one of the pre-defined groups should be assigned zero knowledge. This model is just one proposal. Yockey prefers to use another prescription to determine which amino acids are allowed.

Using the online databases has many advantages. The techniques are easy to use, the databases already exist and can be used to analyze many thousands of proteins. Nevertheless, because of the arbitrary nature of the analytical techniques described in this book, some experimental support is required. Fortunately, this support does exist.

Many researchers have mutated functional proteins by substituting different amino acids at specific sites and then screened the mutated proteins for functionality. Unfortunately, these experiments usually concentrate only on the most critical portions of a protein and thus the results cannot be used here for comparison. Only a few researchers have investigated the effects of a single amino acid change at a random position within a protein. From these experiments, it is possible to directly calculate molecular knowledge and validate any analytical technique that relies exclusively on the online databases.

Quote: "There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection...........

...........The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned genes and uses selections or screens to identify functional sequences.......The end result of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function." [3]

This book only uses the first approach. This appendix will compare the results of this approach to a few proteins that have experimental data. The alignment of one such protein is shown on the next page, and its knowledge is calculated with the techniques developed in chapters 4 and 14 (Table A6). The protein is called 3-methyadenine glycosidase (or ADPG for short).

In reference 1, the researchers create random mutations in AAG the gene that encodes ADPG in humans and then analyze the mutants to see which ones are functional. From this analysis, they estimate that a random substitution in ADPG at a random position will destroy the protein's function 34% of the time. Thus, the average number of allowed amino acids at each position is equal to 20 - (20 * .34) = 13.2 amino acids. Note: this is an average for the whole protein. Some positions may tolerate all amino acids and others may only tolerate 1. From this it is quite easy to calculate the knowledge contributed on average by each amino acid.

Molecular knowledge = 3.32 x log (20 possible / 13.2 found) =
= 0.6 bits per amino acid

The number above is calculated empirically from experimental data. So how does it compare to the analytical techniques developed in chapters 4 and 14? From table A6 (sum of columns 8 and 16), the analytical method assigns 140 bits to ADPG. The shortest sequence in the alignment belongs to E. cuniculi, a very primitive parasitic eukaryote, with 208 amino acids. So the molecular knowledge calculated using the analytical technique is 140/208 = 0.67 bits per amino acid. This is similar to the number calculated for G3PD (0.57 bits per amino acid - see chapter 14).

Clustal Alignment of ADGP

```
            rat -------------SKEPVSV-VLPDAEHPAFPGRTRRPGNARAGSQVTGSREVGQMPAPL    46
          mouse MPARGGSARPGRGSLKPVSVTLLPDTEQPPFLGRARRPGNARAGSLVTGYHEVGQMPAPL    60
          human -------------MVTPALQMKKPKQFCRRMGQKKQRP--ARAG-QPHSSSDAAQAPAEQ    44
    puffer_fish -------------------------------MAGRKRRMVTLEESAAPGQHVNKHERKTGL    30
     E_cuniculi ------------------------------------------------------------
mouse-ear-cress ---------------------------MKTPARRSKRVNQEESETNVTTRVVLRTRKTNC    33
          ruler 1.......10........20........30........40........50........60

                                                             :        .  **
            rat SRKIGQKKQQLAQSEQQQTPKEKLSSTPGLLRSIYFSSPEDRPARLGPEYFDQPAVTLAR   106
          mouse SRKIGQKKQRLADSEQQQTPKERLLSTPGLRRSIYFSSPEDHSGRLGPEFFDQPAVTLAR   120
          human PHSSSDAAQAPCPRERCLGP----PTTPGPYRSIYFSSPKGHLTRLGLEFFDQPAVPLAR   100
    puffer_fish NVSDPERKRSFC-CENNQGD-E----T---VRSCYFT-EKTLQGRLGDRFFNQPCTSLAK    80
     E_cuniculi -------------------------------------------MDMG-----IPCSQLAR    12
mouse-ear-cress SKTRAARVRPDYPLTRTTSE--------------------SEMKLMPPEFFQIDALDLAP    73
          ruler ........70........80........90.......100.......110.......120

                 :**:.: *           *.*.***   :* *.*.  *  * *.  ::   *  ***
            rat AFLGQVLVRRLADGTELRGRIVETEAYLGPEDEAAHSRGGRQTPRNRGMFMKPGTLYVYL   166
          mouse AFLGQVLVRRLADGTELRGRIVETEAYLGPEDEAAHSRGGRQTPRNRGMFMKPGTLYVYL   180
          human AFLGQVLVRRLPNGTELRGRIVETEAYLGPEDEAAHSRGGRQTPRNRGMFMKPGTLYVYI   160
    puffer_fish ALLGKVLIRRCADGTELRGRIVETEAYLGGEDRASHSAGGKRTERNTAMFMKPGTIYVYP   140
     E_cuniculi RLLGKMLCRR-IEGRTTKGMIVETEAYLGKEDKACHSYGGRRTERNSAMYMKAGTCYVYR    71
mouse-ear-cress RLLGKFMRRD-----NVVLRITEVEAYR-PNDSACHGRFG-VTPRTAPVFGPGGHAYVYL   126
          ruler .......130.......140.......150.......160.......170.......180

                 ** :  :*: :     ** **:*: .*: *   ::. *        .   ..  : .**
            rat IYGMYFCLNVSS--QGAGACVLLRALEPLEGLETMRQLRNSLRKSTVGRSLKDRELCNGP   224
          mouse IYGMYFCLNVSS--QGAGACVLLRALEPLEGLETMRQLRNSLRKSTVGRSLKDRELCSGP   238
          human IYGMYFCMNISS--QGDGACVLLRALEPLEGLETMRHVRSTLRKGTASRVLKDRELCSGP   218
    puffer_fish IYGIYLCMNVSS--EGEGAAVLLRSLEPLQGQPTMRQLRATRRK-EGSRELKDKELCNGP   197
     E_cuniculi IYGRYECFNISS--VEAGAGVLVRALEPLCGVSEMRERR-------GGR-VKDRDIANGP   121
mouse-ear-cress CYGLHMMLNIVADKEGVGAAVLIRSCSPVSGMETIQERR--------GLKTDKPVLLNGP   178
          ruler .......190.......200.......210.......220.......230.......240

                .*:  *:       .:. :     : *  .      .   :      *:     *  .
            rat SKLCQALARSK-SFDQRDLAQDEAVWLEHGP-LESSSPAVVAAAA--GIGHAG-EWTQKP   279
          mouse SKLCQALAIDK-SFDQRDLAQDDAVWLEHGP-LESSSPAVVVAAARIGIGHAG-EWTQKP   295
          human SKLCQALAINK-SFDQRDLAQDEAVWLERGP-LEPSEPAVVAAAR-VGVGHAG-EWARKP   274
    puffer_fish SKLCQALDIPR-CFDRRDLASDPEVWLEADAKTDSVEAQRIVTAPRVGVESHG-EWAKKP   255
     E_cuniculi SKLCIAMGITRREIDKEWIAGSEKIWLEEGR--EVADPEIVAGRR-IGIRNCG-EWEEKK   177
mouse-ear-cress GKVGQALGLST-EWSHHPLYSPGGLELLDGG----EDVEKVMVGPRVGIDYALPEHVNAL   233
          ruler .......250.......260.......270.......280.......290.......300

                 ** : .   :*   .
            rat LRFYVQGSPWVSVVDRVAEQMYQPQQTACSDXALIVQK   317
          mouse LRFYVQGSPWVSVVDRVAEQMDQPQQTACSEGLLIVQK   333
          human LRFYVRGSPWVSVVDRVAEQDTQA--------------   298
    puffer_fish LRFYLRGHPCVSVVNKEAEKES----------------   277
     E_cuniculi LRFYIRDNEFVSCIRRRELGNRKHGSVQQLP-------   208
mouse-ear-cress WRFAVADTPWISAPKNTLKPL-----------------   254
          ruler .......310.......320.......330........
```

Table A6: Molecular Knowledge in ADPG

pos	1	2	3	4	5	6	bits	Columns:	pos	1	2	3	4	5	6	bits
118	L	L	L	L	L	L	1.8	1 = rat	198	G	G	G	G	G	G	4
119	A	A	A	A	A	A	1.8	2 = mouse	199	A	A	A	A	A	A	1.8
123	L	L	L	L	L	L	1.8	3 = human	201	V	V	V	V	V	V	1.8
124	G	G	G	G	G	G	4	4 = puffer fish	203	L	L	L	L	V	I	1.8
129	R	R	R	R	R	R	2.67	5 = E. cuniculi	204	R	R	R	R	R	R	2.67
141	I	I	I	I	I	I	1.8	6 = mouse-ear	208	P	P	P	P	P	P	4
143	E	E	E	E	E	E	4	cress	209	L	L	L	L	L	V	1.8
145	E	E	E	E	E	E	4		211	G	G	G	G	G	G	4
146	A	A	A	A	A	A	1.8		215	M	M	M	M	M	I	1.8
147	Y	Y	Y	Y	Y	Y	3.37		219	R	R	R	R	R	R	2.67
152	D	D	D	D	D	D	4		236	L	L	L	L	I	L	1.8
154	A	A	A	A	A	A	1.8		239	G	G	G	G	G	G	4
156	H	H	H	H	H	H	2.67		240	P	P	P	P	P	P	4
160	G	G	G	G	G	G	4		242	K	K	K	K	K	K	2.67
163	T	T	T	T	T	T	2.67		243	L	L	L	L	L	V	1.8
165	R	R	R	R	R	R	2.67		246	A	A	A	A	A	A	1.8
169	M	M	M	M	M	V	1.8		247	L	L	L	L	M	L	1.8
170	F	F	F	F	Y	F	3.67		259	L	L	L	L	I	L	1.8
174	G	G	G	G	G	G	4		265	V	V	V	V	I	L	1.8
177	Y	Y	Y	Y	Y	Y	3.67		267	L	L	L	L	L	L	1.8
178	V	V	V	V	V	V	1.8		280	V	V	V	I	V	V	1.8
179	Y	Y	Y	Y	Y	Y	3.67		281	A	V	A	V	A	M	1.8
182	Y	Y	Y	Y	Y	Y	3.67		287	G	G	G	G	G	G	4
183	G	G	G	G	G	G	4		294	E	E	E	E	E	E	4
189	N	N	N	N	N	N	4		310	V	V	V	V	V	I	1.8
190	V	V	I	V	I	I	1.8		311	S	S	S	S	S	S	2.67

Total number of bits = 140 bits (sum of column 8 and 16). The shortest sequence is 208 amino acids. So the molecular knowledge is 0.67 bits per amino acid. In this case, the molecular knowledge calculated by experimental techniques is almost identical to that calculated using the online databases (0.6 vs. 0.67 bits per amino acid). This is a small difference (on average 13.2 amino acids allowed per site vs. 12.5 amino acids).

Not all proteins will yield ~ 0.6 bits per amino acid when analyzed. Highly conserved proteins will be much higher, and poorly conserved proteins will be much lower.

Nevertheless, 0.6 is an excellent starting point for an average sized enzyme. Such enzymes are more strongly conserved than proteins like hemoglobin, and they are much more variable than proteins like insulin. Experimental results from T4 lysozyme [5] suggest that for this protein only 16% of the amino acid substitutions are deleterious. A similar study for LacI in E. Coli suggests that 34% of the amino acid substitutions deactivate the protein.[3,1,4] Since this is the same deactivation frequency that was observed in AAG, the experimental knowledge is the same or 0.6 bits per amino acid. Does the analytical technique also track the experimetal data for LacI?

From work not shown here, but available on the book's companion web site (theory-of-evolution.net), the analytical technique assigns 0.42 bits per amino acid (so on average each site tolerates 15 amino acids). Yet the experimental data suggests 0.6 bits per amino acid. So in the case, the knowledge assigned by the analytical method is less than that calculated by the experimental method. Since the experimental data is something that can be tested, it should be used as the baseline for molecular knowledge. In other words, the actual knowledge is determined by experiments that randomly mutate amino acids in the protein and monitor protein functionality. Both the LacI and AAG experimental evidence suggests that figure 4.9 in chapter 4 is accurate. That is the analytical approach will yield a value of molecular knowledge that is close to the actual knowledge. Sometimes it may be a little high and sometimes it will be low, but it should always be close.

The real uncertainty with calculating molecular knowledge lies with determining the allowed amino acids. Even the experiments that create random mutations throughout the protein suffer from lack of certainty. In the LacI experiment, only 13 substitutions at each site were analyzed. Furthermore, all of these experiments depend on how sensitive the screens are at detecting barely functional proteins. Due to this uncertainty, the experimental approach may not always be the best way to calculate knowledge; therefore, the next section will compare the group assignments developed in chapter 4 to another analytical method called SIFT (sorts intolerant from tolerant substitutions).

This online computer program uses a complex algorithm to analyze columns in multiple sequence alignments. The program then predicts which amino acids are allowed at each position. Unlike the simple assignment rules developed in chapter 4, SIFT allows groups to overlap. In theory, it should do a better job determining which amino acids are allowed at each site, and it almost certainly does so. The program was developed to help researchers predict which amino acids have harmful effects at each site. SIFT can also predict molecular knowledge.

For comparison, the alignment for ADPG (page 277) was uploaded into the SIFT online program. The program returned tolerant and intolerant amino acids for each site. Based on the alignment above, SIFT indicates that on average 10.73 amino acids all allowed at each site, and this corresponds to 0.9 bits per amino acid. Because E. cuniculi's version of this gene is so short, only the last 220 sites were included. The author's of SIFT did there own evaluation against the LacI protein and found that SIFT predicts 37% of all mutations are deleterious. This corresponds to 0.67 bits per amino acid. From the calculations available on the companion web site, this book's technique only assigns 0.42 bits. So SIFT tends to assign more molecular knowledge to each site than this book's technique. This implies that the techniques developed in this book may underestimate the difficulties associated with creating new proteins.

References:

1) Guo, Choe, Loeb, Protein Tolerance to random Amino Acid Change, PNAS, 2004.

2) Markiewicz, Kleina, Cruz, Ehret, Miller, Genetic Studies of the Lac Repressor, Journal of Molecular biology, 1994.

3) Bowie, Olson, Lim, Sauer, "Deciphering the message in protein sequences: tolerance to amino acid substitutions," Science, March 1990.

4) Saunders, Baker, Evaluation of Structural and Evolutionary Contributions to deleterious Mutation Prediction, Journal of Molecular Biology, 2002.

5) Pottete, Hardy, Genetic analysis of Bacteriophage T4 Lysozyme Structure and Function, Bacteriology 1994.

6) Henikoff, NG, Predicting deleterious amino acid substitutions, Genome research, 2001.

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.

```
                                                        : ***  . :.   : :* .    *    *  *   :  : *  .  *  *       :*:               .:   .                          :
tr|Q7AH57|Q7AH57_ECO57 ----------------------------MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQSLLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVSGL   122
tr|Q6D735|Q6D735_ERWCT --------------------------MKQKPITLHDVAEYAGVSYQTVSRVLNQAPHVSSRTRNKVEQAMAALNYTPNRVAQQLAGKTITTLGLVTTDLSLHAPSQIAAAIKSTASQLGINIVMSMLSSPDVNTCNNAVNDLLSQRVSGV   124
tr|Q666Z9|Q666Z9_YERPS --------------------------MKSKSTTLEDVARHAGVSYQTVSRVLNKSAKVSEATRRKVEQSIELLRYVPNRLAQQLVGKQSMTMGLVTTSLALHAPSQVAAAIKRYAHIEGYQVLISMIDESVNQSIQHSINDLKSQWVGKV   124
tr|Q8ZHP1|Q8ZHP1_YERPE --------------------------MKSKSTTLEDVARHAGVSYQTVSRVLNKSAKVSEATRRKVEQSIELLRYVPNRLAQQLVGKQSMTMGLVTTSLALHAPSQVAAAIKRYAHIEGYQVLISMIDESVNQSIQHSINDLKSQWVGKV   124
tr|Q8CZW9|Q8CZW9_YERPE -MDSGDWVFHLCPCYNQGITLRGSQTMKSKSTTLEDVARHAGVSYQTVSRVLNKSAKVSEATRRKVEQSIELLRYVPNRLAQQLVGKQSMTMGLVTTSLALHAPSQVAAAIKRYAHIEGYQVLISMIDESVNQSIQHSINDLKSQWVGKV   149
  sp|P06220|LACI_KLEPN --------------------------MPRRTATLEDVARRGRVPADGLRRVLNRPEVVSARTREQVIRAMQALHYVPNRSAQLLAGKAAPSIGLITASVTLHAPSQIAAAIKSHASLHQLEVAIAMPAQADFVALQARLDELRAQHIRGV   124
tr|Q4NKP0|Q4NKP0_9MICC ------MRCRLAHNSFMAYSS------ERRLPRLEDVAERAGVSHQTVSRVINNHPNVSKATRERVEAAIAELGYRRNTAARSLVTRRSHTIGVLASELSQYGPANTLLGVEQAARDAGYFVSIAAIREVGAESIADAARHFMDQGVDGI   138
tr|Q4NCE5|Q4NCE5_9MICC ---MRGTAIPREANERLPDTS------LRRKPTINDVAQIAGVSFGTVSRVLNNAPDVSASTRERVLQIVKDIGYRRNRAATALVTSRSTSIGILSDGSPRFGPVGTLMALENIARKKGYATTVISVEKPYEESVQAALDTLDDTGVGGI   141
tr|Q40VM5|Q40VM5_KINRA --------MSAESAPRGEAPEA-----EGRLPVLADVAQLAGVSQQTVSRVVRGSPSVARRTRERVEAAIAELGYRPNIAARTLVTRRSHRIGVVAADTSLYGAATTLAGIQEAARAEGYAVSLVMLPDLSAAGISTALEELRAQYVDGA   137
tr|Q82PA2|Q82PA2_STRAW ---MNTADSGSTSVPRALTRMDEQSARRRQSASMADVARLAGVSSQTVSRVSNGYAGVTEETRQEVLAAMKELGYRPNSAARALKRGEFRTIGVITFNLSTTGNMRTLEAIATSAAQEGYAVTLLPVAVPTQDEVHGAFSRLGELAVDAV   147
tr|Q9L148|Q9L148_STRCO ---MSTAEEVPTSAPAPPRRG------RDRTASMADVARLAGVSSQTVSRVSNGYAGVTEETRQQVLAAMKELGYRPNSAARALKRGEFRTIGVITFTLATTGNVRTLEAIATSAASEGYAVTLLPVAVPTQDEVRGAFSRLEELAVDAV   141
tr|Q4R0J8|Q4R0J8_STRCX ----MTRGSADDSAPRGRSRRN----FAGSRPVMDDVARLAGVSKQTVSRVLNDHPSVRAETRETVLSAMRTLGYRPSRSARSLASGRTRMVGVISFDAARYGPASILTAINTAAQEAGYLLSSIALDTAEPDTVVEAVNRLSAEGADGV   142
tr|Q4NCF2|Q4NCF2_9MICC MSTTPAPSNARNSNGRGERLAS-----RSKAPRMSDVARLAGVSQQTVSRVVRGATNVDPDIRERVEHAIAQLRYRRNPAAAALASNRTMTIGVVSFELSVLGPTVALYGISEEARKHGYATRLVTLASLERRDIRAAFESINSDTVDGV   145
tr|Q4ZUH5|Q4ZUH5_PSEU2 ------------------------------MATIKDVAALAGISYTTVSHVLNNTRPVSEPVRLKVEAAIAQLDYVPSAVARSLKAKTTSTIGLLIPNGMNPYFAELARGIEDYCERNGFCVILCNSDDN-PEKQRGYLRVLLEKRVDGL   119
                 ruler 1.......10........20........30........40........50........60........70........80........90.......100.......110.......120.......130.......140.......150

                       :                                           .      :*   **  .  . *.        *  .:   *               :: .  .      :        *:....*  * * : *  :         *
tr|Q7AH57|Q7AH57_ECO57 IINYPLDDQDAIAVEAACANVPALFLD-VSDQTPINSIIFSHEDGTRLGVEHLVALGHQQIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREG--DWSAMSGFQQTMQMLNEGIVPTAMLVANDQMALGAMRAITESGLRVGADIS   269
tr|Q6D735|Q6D735_ERWCT IVNVPLSTDEVAHISQTCADTPVLFMD-ADPHTDILNVMFDPDHGARLAIAHLVELGHQHIALLTGPMTSISARLRYEGWLAELEKQRLSPVSVLHG--DWSAASGYHQTLALLGQSLRISAIVVANDQMALGVLRALHEYGLRVPEQMS   271
tr|Q666Z9|Q666Z9_YERPS IINVPLETAVAEKIAADNDDVTCLFLD-VDPYSSVFNVSFNPADGTRASVKYLYELGHREIALLAGPKQSVSANLRLKSWLETLADYGLSPVSVIHG--NWDAQSGYSGALQMLRETPQFSAVLVANDQMALGVLSAFHQNQLAVPGEKS   271
tr|Q8ZHP1|Q8ZHP1_YERPE IINVPLETAVAEKIAADNDDVTCLFLD-VDPYSSVFNVSFNPADGTRASVKYLYELGHREIALLAGPKQSVSANLRLKSWLETLADYGLSPVSVIHG--NWDAQSGYSGALQMLRETPQFSAVLVANDQMALGVLSAFHQNQLAVPGEKS   271
tr|Q8CZW9|Q8CZW9_YERPE IINVPLETAVAEKIAADNDDVTCLFLD-VDPYSSVFNVSFNPADGTRASVKYLYELGHREIALLAGPKQSVSANLRLKSWLETLADYGLSPVSVIHG--NWDAQSGYSGALQMLRETPQFSAVLVANDQMALGVLSAFHQNQLAVPGEKS   296
  sp|P06220|LACI_KLEPN IVSLPLESATAERLVQDNPDMACLFLD-VSPEADVCCVRFDHRDGCGACVRHLWEMGHREFGLLAGPESSVSARLRLASWREALHSLNIARSTTVFG--DWSAASGWQKTFELLHLQPRISAIVVANDQMALGVLSALAQLNRSGSQAVS   271
tr|Q4NKP0|Q4NKP0_9MICC VALVPHAGTLQALEQLNLDVPVVVVGS--SGAGRFSGAMVDQKMGARLAVEHLIERGHKRIGHISGPRDWIDAAERAEGWREALAAEGLPADLLVEG--DWSAGSGYGIGQKLAVER-NATAVFVGNDQMALGVLRAFSEAGVRVPRDVS   283
tr|Q4NCE5|Q4NCE5_9MICC IVIAPLVDMAAAVWNASCRVPVEMIAAGASSTPNVFTYSENQELGARLATQHLIDLGHTDIAHFAGSMDWFDGRVRKRGWEAALRDAGLAPGLCIEG--DWSPKWAYETGLQLVREGRIPQAIFAASDHTALGLIRAFAENGVRVPDDVS   289
tr|Q40VM5|Q40VM5_KINRA IAVVPQDAAQDAVRAADAAFPCVLAPG--LDGAGVPDAYWAEVAAAREATHHLLDLGHPTVHHVAGPPDWAESRARATGWRTALVERLRVVPRPVRG--DWSAQSGYTAVGDLP-LE-EVSALFVGNDQMAAGVLRALADRGRRVPDEVS   281
tr|Q82PA2|Q82PA2_STRAW IIIMEVHLLDAATISLPPHVQVAVADS--DAGDRYTVVDTDQAGGARAAVRHLLELGHRTIWHLAGPEESYAAQRRADAWRAELIEAGRVPPAVVRG--DWSAESGYQAGLRIADEE-DCTAVFAANDQMALGLLRALHERGRRIPEDVS   292
tr|Q9L148|Q9L148_STRCO IVIMEVHLLDAASLKLPPHVQVVVVDS--DAGDTYTVVDTDQAGGSRAAVRHLLDLGHDTVWHLGGPEGSFAAQRRTDAWRRTLTEAGRTPPPLVRG--DWSAESGYRAGLELAARE-DCTAVFAANDQMALGLLRALHERGRRVPEDIS   286
tr|Q4R0J8|Q4R0J8_STRCX IAIAPQLWVGKALADTRLDTPLVVLEN--GVDADTQQVTGDSRAGARKAVGYLLDLGHRTVWHVAGPTGWTSADHRLGSWQACLEAAGAAVPTPLIG--DWSADSGYELGRRLARRT-DVTAVFASNDQMALGLLHALHEAGRSVPGDIS   287
tr|Q4NCF2|Q4NCF2_9MICC IVLAPLLEAIAVLAGIDIGVPVVTFQQ--GSHPSPTSVSVDEVRGARMVVRHLLDLGHKTVWHVQGPPGWMATSARVQGWAAELGASGRFVPTPIATT-DWSAAEGYRVGRELAARE-DVTAVFAANDPFALGVIKALDEAGRDVPGDVS   291
tr|Q4ZUH5|Q4ZUH5_PSEU2 IVSSVGGDAGIAGGLAEVRTPLVIVDR-ELDGIDADMVRIDHEQGAYLATRHLLDLGHRAIACIGGPFHSTVAEMRLAGYRRALQQASVPVRAEWALHSEFTSSAGHEAACQLLAAH-PPTAIFAGNDVIAIGVLRAAAERLIRVPQDLS   267
                 ruler .......160.......170.......180.......190.......200.......210.......220.......230.......240.......250.......260.......270.......280.......290.......300

                       * *:**   :    *.***:  .:   .      ::                   :: * ::
tr|Q7AH57|Q7AH57_ECO57 VVGYDDTEDSSCYIPPLTTIKQDFRLLGQTSVDRLLQLSQ--GQAVKGNQLLPVSLVKRKTTLAPKTQTASPRALADSLMQLARQVSRLESGQ   360
tr|Q6D735|Q6D735_ERWCT VIGFDDTQDSAYYTPPLTTIRQDFRLLGKEGVTRLLATLR--DPTQATSLLLPTELIVRHSTAIHSSTHASLQELTKNLLDITRQLQRL----   358
tr|Q666Z9|Q666Z9_YERPS VIGYDDTYESSFFYPALTTVSLDLDLQGKEAVRRMLSASD--NESMRSSSILPAKLVVRNSTGPQAKEHRDLQKIAEELRLIAQRLAS-----   357
tr|Q8ZHP1|Q8ZHP1_YERPE VIGYDDTYESSFFYPALTTVSLDLDLQGKEAVRRMLSASD--NESMRSSSILPAKLVVRNSTGPQAKEHRDLQKIAEELRLIAQRLAS-----   357
tr|Q8CZW9|Q8CZW9_YERPE VIGYDDTYESSFFYPALTTVSLDLDLQGKEAVRRMLSASD--NESMRSSSILPAKLVVRNSTGPQAKEHRDLQKIAEELRLIAQRLAS-----   382
  sp|P06220|LACI_KLEPN VTGYDDTADSLYFQPPLTTVAQDFDLLGKRAVERLIALMA--APQLRIRELLPTRLIVRQSAWPVAAAEDRQQTLAQLKALVEKL--------   354
tr|Q4NKP0|Q4NKP0_9MICC VVGFDDQPEAGYFTPPLTTVRQDFEELGRRCVDLMLRLLE--DGAAAGTTVVEPRIVVRASTARPDASPR-----------------------   351
tr|Q4NCE5|Q4NCE5_9MICC VVGFDDIEGSDFFLPPLTTVRQDFTALALMSTEVLLGAME--G-RDVDRTPITPTLVVRNSSTRATPTARHKKTPEAPPNKA-----------   368
tr|Q40VM5|Q40VM5_KINRA VVGFDDVPEARFYFPPLTTVHQDFTEIGWRCVRVLLGRLH--DRVVEPGPVTS-TLVVRSSTAPPPAPRVAGRPGDQALRG------------   359
tr|Q82PA2|Q82PA2_STRAW VIGFDDIPEAGSFLPPLTTVHQDFAEVGRLCVQGVLRRMR-HDAAEHGTTLVPTKLVVRDSTAPPPSR-------------------------   359
tr|Q9L148|Q9L148_STRCO VIGFDDIAEAGSFLPPLTTVHQDFAEVGRLCVRAVLRKMR-AAGPEQGTTLVPTRLVTRASTAPPPR--------------------------   352
tr|Q4R0J8|Q4R0J8_STRCX VVGYDDIPEAAHFLPPLTTVRTDFAEIGTRSLRLLLDRLDGTEEPARTRSLVPVDLVIRASSGPPPTR-------------------------   355
tr|Q4NCF2|Q4NCF2_9MICC VVGFDDVKEAPFFRPALTTVRLDFEAIGHIAVGRILAMIK--DEAEGDIPLLEPRLVLRDTTAPPPSQR------------------------   358
tr|Q4ZUH5|Q4ZUH5_PSEU2 VIGFDDIQMSRYVYPALTTVGQSIMQLGETAAEMLLSRIATPHALPAEKRLVTPSVVVRESTAAPNTLSNE----------------------   338
                 ruler .......310.......320.......330.......340.......350.......360.......370.......380.......390...
```

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	V	L	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	P	R	R	P	P	P	P	R	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	L	V	I	V	V	V	V	V	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	I	V	L	I	I	I	I	I	I	1.8
150	L	V	V	V	V	V	I	I	A	V	V	V	V	L	1.8
151	I	I	I	I	I	I	V	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	E	D	D	E	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	l	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	l	f	H	Y	E	20
8D	D																				1
9V	V																				1
10A	A																				1
11E	l	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	l	f	R	H	Y	20
13A	A																				1
14G	G																				1
15V	V																				1
16S	S																				1
17Y	t	r	n	s	v	a	l	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	l	g	t	d	a	s	e	r	N	Q	K						15
27A	w	c	m	i	f	p	y	H	v	l	q	r	d	n	k	g	e	t	S	A	20
28S	r	l	q	n	d	k	v	e	t	g	S	P	A								13
29H	m	f	i	v	y	p	l	t	a	q	g	s	e	d	r	H	N	K			18
30V	V																				1
31S	S																				1
32A	m	h	i	p	v	l	g	n	r	t	q	d	S	A	K	E					16
33K	m	y	i	h	v	p	l	g	d	n	t	s	q	e	A	R	K				17
34T	T																				1
35R	R																				1
36E	y	m	i	h	v	p	l	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	l	n	p	v	r	t	d	Q	g	k	s	e	A		19
41A	t	g	S	A																	4
42M	a	f	v	l	M	I															6
43A	h	i	l	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	l	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	l	I	V		18

49P	f m y i h v d l 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 I 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 l 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	l 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 l 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 I A v f l Y	20

93S	w c m p i h F v g r Q l d y N t k a e S	20
94V	l 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	l 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 I	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 l 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	l V I	3
124I	f m t l 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 I 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 I	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 I V	5
160I	w c h d p g n q r e y k M f S a t l v I	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	fm y i v p l g n t d s a H Q k R E	18
182I	I	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 I 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 ym h i v p l g r n Q D a T k S e	19
191S	m cW h i q r e p v k l d f y n g a t S	20
192V	l I V	3
193S	h r p q k a e g t n D S	12
194A	A	1
195R	cm 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 km 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 dm 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 cm 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 fm h i p v l g n t R q D s A K e	19
208N	w cm p i v h f g l t r Q N Y s a d k E	20
209Q	fm i y h v p l t d n s a G e Q R k	18
210I	m v I L	4
211Q	c fm 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 sm f t a l I V	12
214A	hm 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 tm f a L V I	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	l 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 I 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 I A y v l	20
240T	n a T S	4
241A	A	1
242M	l 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 I 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 I V	18
269S	S	1
270V	V	1
271V	l V I	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 I 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 I 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 I	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 I L	4
324V	l I 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 I 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 I	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 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 differences explain why the techniques used by this book will always assign less molecular knowledge to a protein than SIFT.

Glossary:

Activation Energy - an energy barrier that must be crossed before chemicals can react.
Adenine - one of the bases in DNA and RNA. It is also one of the components in ATP, ADP and AMP.
ADP - adenine diphosphate. ADP only has one high energy bond.
Amino Acid - the building blocks of proteins. Life uses 20 amino acids to build proteins. Amino acids have two sticky ends that can be joined together to form long chains.
AMP - adenine monophosphate. AMP has no high energy bonds. It is created in the cell from ATP when ATP is used to accomplish some task that requires energy.
ATP - adenine triphosphate. This is the energy source used by life. It is made by plants during the process of photosynthesis. It is made by animals as they digest food. ATP has 2 high energy bonds. Life knows how to use the high energy bonds to do work.
ATP synthase - an enzyme that synthesizes ATP.
Axiom - a self evident assumption.
Bit - a unit of information or knowledge.
Carboxylic acid - one of the sticky ends on an amino acid.
Cell - the smallest living unit. Cells are surrounded by a membrane which is composed of lipids and proteins. The chemicals inside cells know how to grow and replicate. They accomplish this by using an abundant energy source to do useful work.
Chemical evolution - the hypothetical process by which the chemicals necessary for life emerged on the primitive earth.
Codon - a grouping of 3 bases in DNA or RNA that specifies a specific amino acid in the final protein.
Common Ancestor - a animal, plant or bacteria whose descendants evolved into two or more species.
Complexity - a non-repetitive pattern.
Condensation agent - a chemical that facilitates the formation of chemical bonds between biological molecules by absorbing water.

C-terminus - one of the sticky ends of an amino acid.

Deoxyribose - the sugar molecule found in DNA.

DNA - stands for deoxyribose nucleic acid. DNA is the molecule that stores all of the knowledge that life needs to grow and replicate. Sections of DNA that contain the knowledge to build proteins are called genes.

Energy - the ability to do work.

Entropy - a measure of uncertainty.

Enzyme - a protein that catalyzes a chemical reaction.

Eukaryote - cells with a defined nucleus. Plants and animal cells are composed of eukaryotic cells.

G3PD - an enzyme used by life to metabolize sugar.

Gene duplication - the process by which existing genes are duplicated.

Gene - a section of DNA that contains the knowledge to build a protein.

Genetic Code - the code that is used to build proteins from the knowledge contained in DNA.

Heat - the flow of energy from hot to cold objects.

Hydrophobic - a molecule that does not like water.

Knowledge - a useful reduction in uncertainty.

Infon - a step in molecular knowledge found by chance.

Information - a reduction in uncertainty. Information is not necessarily useful.

Intelligent design - a methodology that relies on indirect logic to infer the existence of a creator.

Investigator interference - process by which researchers alter the results of their experiments.

Irreducible complexity - a system that requires two or more components to function.

Micro-state - the arrangement of particles in a system.

Molecule - a chemical composed of two or more atoms.

Molecular Knowledge - the minimum amount of information necessary to enable a chemical or group of chemicals to accomplish some task or to specify some trait.

mRNA - stands for messenger ribonucleic acid. RNA is an intermediate molecule involved in protein synthesis.

Natural selection - the process by which nature ensures that only optimized genes are passed onto future generations.

N-terminus - one of the sticky ends of an amino acid.

Nucleotide - the building block for DNA and RNA. Each nucleoide contains 1 phosphate, 1 ribose, and one of the 5 bases, adenine, guanine, cytosine, thymine, or uracil.

Peptide - a short chain of amino acids.

Perpetual motion machines - a machine that violates one or more of the laws of physics.

Phosphodiester bond - the high energy bond between phosphate groups in ATP.

Polar - a molecule that likes water.

Poly-nucleotide - a string of nucleotides.

Pre-RNA - hypothesized molecule that proceeded RNA during chemical evolution.

Primordial soup - the hypothetical small pond in which life originated.

Primordial information - the knowledge required to exclude chemicals in the primoridal soup from growing biological molecules.

Prokaryote - cells without a defined nucleus. Bacteria cells are prokaryotic cells.

Protein - a chemical composed of amino acids that is used by life to implement the molecular knowledge found in genes.

Protein domain - a section of a protein that performs some function or specifies some 3-D shape.

Proteinoids - long branched chains of amino acids formed by heating.

Observable axiom - the assumption that man is capable or accurately observing the world around him.

Order - a repetitive pattern.

Oxidation - a chemical reaction in which electrons are transferred from one atom to another.

Quantum mechanics - the physics that describes small particles.

Rasmol - a free computer program that allows users to view molecules like proteins and DNA.

Relative entropy - a measure of uncertainty.

Ribose - the sugar molecule found in RNA.

Ribozyme -an RNA molecule that also functions as an enzyme.

Ribosome - a complex of RNA and proteins. Proteins are built by the process of translation at ribosomes.

RNA - RNA contains molecular knowledge and can sometimes implement molecular knowledge. Messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) help cells use the knowledge contained in DNA to build proteins.

Second law of thermodynamics - states that the entropy of the universe increases with time. This happens because particles always try to find their most probable distribution. This is the distribution that maximizes the number of available micro-states.

Shannon entropy - the average uncertainty per symbol.

Species - a group of interbreeding animals.

Specified Complexity - any outcome that is not ordered and is predicted in advance.

Swiss Prot - an online Database that contains the amino acid sequence of most proteins.

Thermal proteins - long branched chains of amino acids formed by heating.

Thermodynamics - the field of physics that deals with heat, energy and work. Chemical thermodynamics is a subset of this discipline that deals with chemicals and how they interact.

Transcription - the process that writes DNA into mRNA.

Translation - the process that uses mRNA to create proteins.

Thymine - one of the bases found in DNA.

tRNA - stands for transfer ribonucleic acid. Transfer RNA brings amino acids to ribosomes for protein synthesis.

Uracil - one of the bases found in RNA.

This book would not have been possible without numerous online programs and resources.

The Protein Database:
www.rcsb.org/pdb/ The protein database that houses the 3-D structure of many proteins.

Rasmol:
www.umass.edu/microbio/rasmol/distrib/rasman.htm Displays 3-D structures of proteins in the protein database.

Swiss Prot:
www.us.expasy.org/sprot/ Contains the amino acid sequence of most proteins.

Consurf:
www.consurf.tau.ac.il/ Colors 3-D displays by amino acid conservation.

Glaser F./ Pupko T., Paz I., Bell R.E., Becher D., Martz E., Ben-Tal N., Consurf: Identification of Functional Regions in Proteins by Surface Mapping of Phylogentic Information, Bioinformatics, Vol. 19, no, 1, 2003 pp163-164.

Clustal:
http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html An easy to use windows interface for the Clustal alignment program.

SIFT:
http://blocks.fhcrc.org/sift/SIFT_help.html

This Book's Companion Website:
http://www.theory-of-evolution.net

Subject Index:

Activation Energy 123-124, 130-133
Activated Nucleotides 173
Adenine Synthesis 207-215
Amino Acids 149-155
ATP 94, 217, 223
  Synthesis 234-236
Calculated Molecular Knowledge 83
Cambrian Explosion 14
Chemical Bonds 108-109
Chemical Equilibrium 124
Chemical Evolution 8
Chemical Kinetics 122
Chemical Oscillators 139-140
Common Ancestor 68, 214-216
Complexity 31
Condensation Agents 167, 171
Cytosine 169
Darwinian Evolution 36
DNA 46-47, 142-145
  Replication 47-48
Energy 8, 111
Enzymes 129-135
Evolution by Design 36
Evolution of Knowledge 93
Eukaryotes 60
Gene Duplication 3
Genetic Code 49
Information 17, 18, 19, 24, 25, 31, 38-39
Infon 29
Insulin (information and Knowledge) 63-85
Investigator Interference 161, 174-177
Irreducible Complexity 200-202
Matter 111
Meteorites 90
Micro-states 114-120
Miller’s Experiment 162-163
Molecular Knowledge 4, 5, 11, 18, 28-39, 83, 101, 102, 212-214, 226-232
Molecular Information 4
Molecular Theory of Evolution 2
Mutation 2, 49
Natural Selection 4,5 250-254
Naturalistic Axiom 1, 267
Non-equilibrium Thermodynamics 137
Observable Axiom 267
Open Systems 127
Order 31
Origin of Life 10, 214, 248
Oxidation 220-222
Peptides
  Short 165
  Long 166
Perpetual Motion 12, 191, 194

Prebiotic Synthesis
  Nucleotides168-170
Pre-RNA 171
Primordial Soup 8, 95-97, 196
Prokaryotes 60
Proteins
  First 92, 98
  Structure 57-59
  Translation 40-41, 52-56
  Thermal 164
  Self Replication 181-183
Proton Gradient 222
Quantum Mechanics 112-116
Relative Entropy 261-264, 272
RNA 144-147
  Self Replication 10,12,13, 184-195
Second Law 116, 117, 118, 129-136
Shannon Entropy 27, 259-260
Specified Complexity 32
Tree of Life 205, 218
Time 239, 244, 247

Author Index:

Bartel 187-188
Behe 200-202
Brillouin 27
Crick 2
Dawkins 7
Darwin 2, 35-37, 85, 252
Dembski 31-32
Ferris 8
Fox 9, 164
Joyce 8, 185-186
Kauffman 139-140
Lee 181
Mayr 215
Meyer 14
Miller 8, 162-163
Ohno 3
Orgel 8, 11,185-186
Schneider 262-274
Shapiro 8, 169-170
Thaxton 8, 161, 169-170
Watson 2
Woese 214-215