Chapter 4: Information and Knowledge in the Protein Insulin

This chapter will calculate the information and molecular knowledge in a real protein. The techniques discussed in this chapter to calculate knowledge are somewhat arbitrary because they rely on both math and human insight. Furthermore, the techniques used here to calculate information differ slightly from those used by other authors. For justification, interested readers are referred to appendixes 2, 4, and 5. Appendix 6 in particular will provide experimental support for the analytical techniques developed in this book. While the analytical techniques are arbitrary, they do seem to work quite well.

For this chapter, a small protein is desirable. Insulin meets this criteria. Insulin is a special kind of protein known as a hormone. When insulin is released into the blood stream, it signals cells to absorb sugar. The actual hormone consists of two short chains, A and B. This chapter will calculate the information and knowledge in both chains. The A chain contains 21 amino acids and the B chain contains 30. The B chain will be considered first.

The most common sequence for chain B in mammals is as follows:

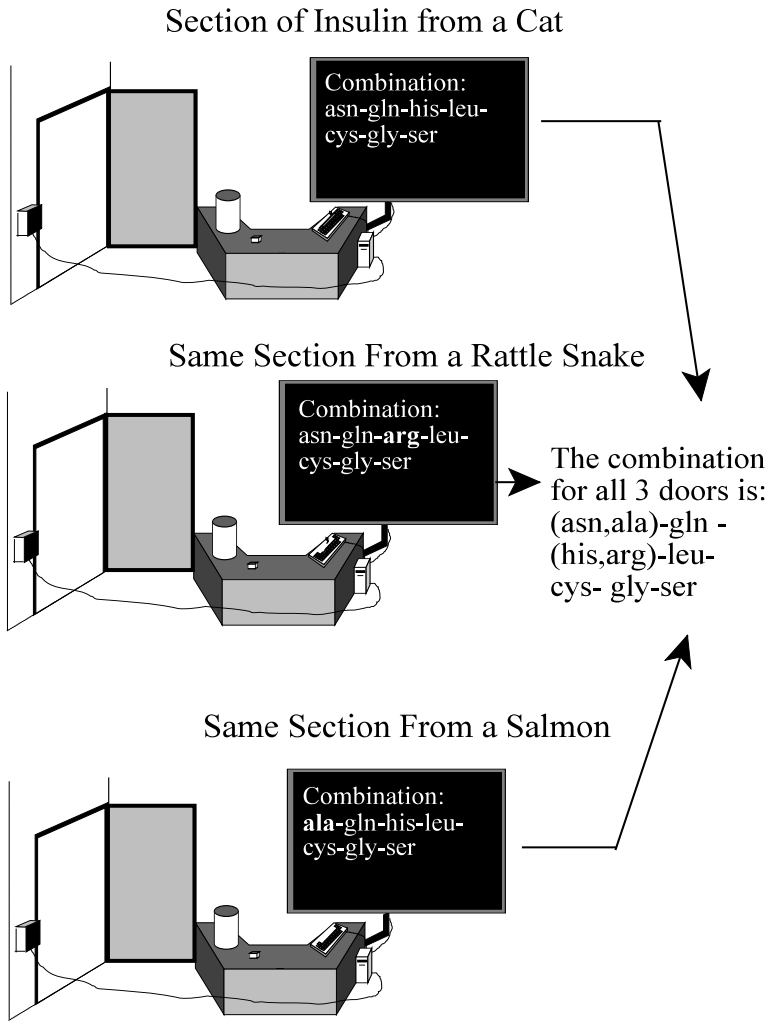
phe-val-asn-gln-his-leu-cys-gly-ser-his-leu-val-asp-ala-leu-
tyr-leu-val-cys-gly-glu-arg-gly-phe-phe-tyr-thr-pro-lys-ala

At several positions in this chain more than one amino acid is allowed. The number of allowed amino acids at each position is determined by comparing the insulin found in man to that in pigs, cats, dogs, fish, frogs and snakes. The next section will illustrate this technique.

Determining Allowed Amino Acids

Because insulin exists, the doors are open in figure 4.1, and the scientist has left. In all three cases, the door's combination is 7 amino acids long. The combination that opened each door is shown on the screen. The combinations are very similar, but there are differences.

Figure 4.1: A Section of Insulin in Cat, Snake and Fish



These 7 amino acid combinations corresponds to a section of the real insulin protein in cat, snake and fish. The differences are important in determining the allowed amino acids. The first amino acid is asparagine (asn) in rattlesnakes and cats, but it is alanine in salmon; thus, both ala and asn are acceptable at position number 1. At position 3, histidine (his) is found in both cats and fish, but arginine (arg) is found in snakes; thus, both his and arg are acceptable at position number 3. In figure 4.1, positions where more than one amino acid is allowed are represented by placing the allowed amino acids in parenthesis; therefore, position 1 is represented by (asn, ala) which means that either amino acid is fine at this position. From a comparison just like this, it is possible to determine which amino acids are allowed at every position in the A and B chains of insulin, and this will determine the amount of information contained in insulin today.

How Much Information Opens the Door

In figure 4.1, the composite combination, (asn, ala)-gln-(his, arg)-leu-cys-gly-ser, is assumed to be functional. In other words, this combination will open all three doors in figure 4.1. To compute the information in this composite combination, the odds of each amino acid arising by chance at each position must be known.

Table 4.1 - Odds of Each Amino Acid Arising by Chance

Number of Codons	Odds	Amino Acids
6	6 in 64	ser, arg, leu
4	4 in 64	val, pro, thr, ala, gly
3	3 in 64	ile
2	2 in 64	phe, tyr, his, gln, asn, lys, asp, glu, cys
1	1 in 64	met, trp

The odds of each amino acid arising by chance are listed in table 4.1. For example, the amino acid, serine (ser), will arise by chance 6 times in every 64 tries. The same is true for leucine(leu) and arginine(arg). The amino acid, methionine (met), will only arise by chance 1 time in 64 tries. Table 4.1 was created from table 3.2 by assuming that all mutations are random.

Table 4.1 will now be used to calculate the information in figure 4.1. In figure 4.1, the required combination is as follows:

pos1	pos2	pos3	pos4	pos5	pos6	pos7
asn	gln	his	leu	cys	gly	ser
ala		arg				

Position 1: asn (odds = 2 in 64) and ala (odds = 4 in 64) are both allowed. The odds of seeing either an asn or ala are found by simple addition. 2/64+4/64=6/64. In other words, the odds of an ala or asn arising by chance at position one are 6 times in 64 tries. Using the second equation presented in chapter 1, Information = 3.32xlog(64/6) = 3.4 bits. Position 1 contains 3.4 bits of information.

Position 2: gln has a 2 in 64 chance of arising by chance. This is equivalent to 1 chance in 32 tries. So the information is easy to find: $2^{(information)}$ = 32/1. Since 2^5 =32, position 2 must contain 5 bits of information. Notice that logarithms can also be used. Information = 3.32 x log(32/1) = 5 bits. Both equations give the same result.

Position 3: The odds for histidine (his) are 2 in 64. The odds for arginine (arg) are 6 in 64. The sum determines the odds that one of these will be present. The sum is 8 in 64 which is equivalent to 1 in 8. The information is $2^{(information)}$ = 8/1. Since 2^3=8, position three contributes 3 bits of information.

Position 4: Leucine has a 6 in 64 chance of arising by chance. These odds are the same as position 1. So position 4 also contributes 3.4 bits.

Position 5: cysteine (cys) has a 2 in 64 chance of arising by chance. These odds are the same as those calculated for position 2. So position 5 contributes 5 bits.

Position 6: glycine (gly) has a 4 in 64 chance of arising by chance. This is equivalent to 1 in 16. So the information at position 6 is $2^{(\text{information})} = 16/1$. Since $2^4=16$, position 6 contributes 4 bits of information.

Position 7: serine has a 6 in 64 chance of arising by chance. These odds are the same as those for position 4. So position 7 contributes 3.4 bits of information.

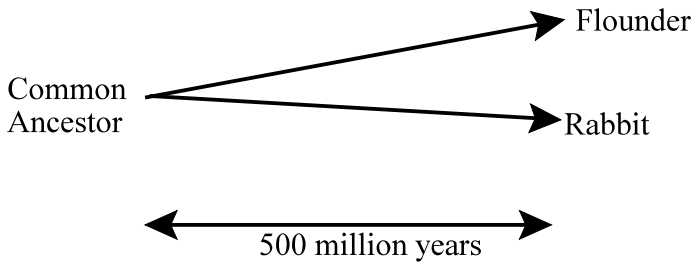
The total information required to open the door is the sum of the information found at each position or 3.4+5+3+3.4+5+4+3.4 = 27.2 bits. The odds that this door can be opened by chance are given by1 in $2^{27.2}$ or 1 in 154 million.

But the above calculation is wrong! The previous paragraph assigns a probability to a protein evolving based on its information content today. It completely ignores the ability of natural selection to guide evolution. While the information calculated is correct, 27.2 bits, this information has no relationship to the probability of any protein evolving. In figure 4.1, only the last door for each of the three species is shown. There may be many doors leading up to these doors, and the odds for success may be quite good. Information should never be related to a probability when natural selection is involved.

Why Does this Work? Common Ancestors

The evolution of all modern proteins can be traced back to a common ancestor. Figure 4.2 illustrates this concept. Around 500 million years ago suppose that there was a common ancestor for flounders and rabbits. Some of the descendants of this ancestor evolved into rabbits. Others evolved into flounders. Today the DNA of this common ancestor is not available, but the DNA of flounders and rabbits is certainly available. When the DNA of a flounder and a rabbit are compared, most of the information found in their DNA is the same. The insulin found in a flounder is very similar to that found in a rabbit, but there are differences because the two species have had 500 million years to accumulate changes independently.

Figure 4.2: Common Ancestors



If a mutation modifies the insulin amino acid sequence, several fates exist for the modified protein.

- If the modified protein is better than the original, natural selection may encourage it to spread through the population. With time the new protein may be fixed in the population. This means that every member of the population possesses the modified protein.

- If the modified protein provides no selective advantage, it may still be fixed in the population. As long as the modified protein is as good as the original, but no better, the probability of fixation is equal to the rate of change.[3] Any protein that meets these criteria is termed neutral. So if a specific amino acid in insulin mutates every 100 million years, then a modified insulin with the changed amino acid is expected to be fixed in the population every 100 million years.

- If the mutation is slightly harmful, natural selection will most likely eliminate it from the population but not always.[3]

Assume for a minute that the amino acid sequence of insulin is not important and that almost any protein composed of 50 amino acids or more can signal cells to absorb sugar. In other words, the insulin hormone contains almost no useful information. If this assumption is true, then one would expect the insulin amino acid sequence in fish and in mammals to be completely different. The sequences have had 500 million years to change independently.

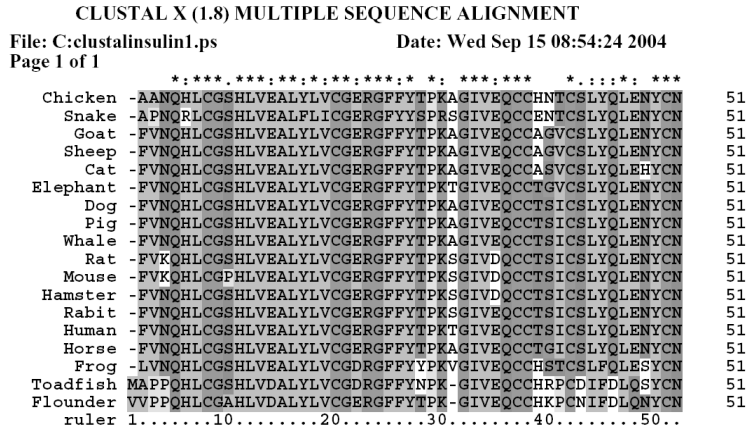
Analysis of insulin in fish and in mammals has revealed that this is not the case. Many of the amino acids are the same or have similar chemical properties. These amino acids are said to be conserved.

Conserved amino acids are a measure of information. To accurately measure this information, a comparison of many diverse species is required. The more diverse the species the better. The technique works best for proteins that are found in all of the kingdoms of life.

Total Information in Insulin A and B Chains

The techniques used in the previous section can easily be extended to calculate the total information in insulin. Figure 4.3 was generated using a software package designed to align the amino acid sequences of similar proteins, Clustal X. Instead of using the three letter abbreviation for each amino acid, the single letter abbreviation is used to conserve space. The letters across the first row are the amino acid sequence of insulin in chickens, the second row snakes, and the last row flounders. The columns are aligned by the Clustal X program so that similar amino acids appear in the same column. The dashes represent gaps inserted by the Clustal program to align the sequences.

Figure 4.3: The Amino Acid Sequence of Insulin



Single letter amino acid abbreviations used in figure 4.3:

A = Ala	C = Cys	D = Asp	E = Glu
F = Phe	G = Gly	H = His	I = Ile
K = Lys	L = Leu	M = Met	N = Asn
P = Pro	Q = Gln	R = Arg	S = Ser
T = Thr	V = Val	W = Trp	Y = Tyr

Insulin is composed of two chains, A and B. Referring to the ruler at the bottom of figure 4.3, the B chain is comprised of amino acids 2-31, and the A chain is comprised of amino acids 32-52. The columns are aligned in such a way to match up the amino acids when they are the same. For example, the last amino acid on each row is always N. This means that N (the amino acid Asparagine or Asn) is conserved. Notice that even in columns where the amino acids differ, the variability is still quite low. Insulin is a highly conserved protein.

Tables 4.2 and 4.3 calculate the information in insulin as it exists today. To help with understanding, a few sample calculations are described first.

Example calculation for table 4.2: refer to figure 4.3, and find the 10^{th} column. Notice that it contains A (alanine) in flounders, S (serine) for most species, and P (proline) in mice. This means that all three of these amino acids are acceptable at this position. The information content is calculated as follows (refer to table 4.1): information = 3.32 x log[64 possible outcomes/ (4+6+4) observed outcomes] = 2.2 bits. Thus, row 10 in table 4.2 is assigned 2.2 bits.

Example calculation for table 4.3: refer to figure 4.3, column 42. This column is always C (cysteine), thus, the information is as follows: information = 3.32 x log [64/2] = 5 bits. The total information for each chain is the sum of the information at each position. The sum of columns 3 and 6 in table 4.3 is 81 bits.

Table 4.2: Information in Insulin (B chain only)

pos	allowed amino acids	bits	pos	allowed amino acids	bits
2	phe, ala, leu, val	2.0	17	phe, tyr	4
3	val, ala, pro	2.4	18	leu	3.4
4	pro, lys, asn	3	19	val, ile	3.2
5	gln	5	20	cys	5
6	his, arg	3	21	gly	4
7	leu	3.4	22	asp, glu	4
8	cys	5	23	arg	3.4
9	gly	4	24	gly	4
10	ala, pro, ser	2.2	25	phe	5
11	his	5	26	phe, tyr	4
12	leu	3.4	27	tyr	5
13	val	4	28	thr, ser, asn	2.4
14	glu, asp	4	29	pro	4
15	ala	4	30	lys, arg	2.7
16	leu	3.4	31	ala, thr, ser, -	0

Total = 108 bits

Table 4.3: Information in Insulin (A chain only)

pos	allowed amino acids	bits	pos	allowed amino acids	bits
32	gly	4	43	ser, asn, asp	2.7
33	ile	4.4	44	leu, ile	2.8
34	val	4	45	phe, tyr	4
35	glu, asp	4	46	gln, asp	4
36	gln	5	47	leu	3.4
37	cys	5	48	glu, gln	4
38	cys	5	49	asn, ser,his	2.7
39	glu, his, thr, ala	2.4	50	tyr	5
40	asn, lys, arg, ser, gly	1.67	51	cys	5
41	pro,thr,ile,val	2.1	52	asn	5
42	cys	5			

Total = 81 bits

The total information today in insulin is the sum of the information found in both chains or 81 +108 = 189 bits. It is important to keep in mind that this number has absolutely nothing to do with the probability of insulin evolving.

Molecular Knowledge of Insulin

Finding the information contained in insulin is straight forward. The math is tedious, but the procedure is at least defined, and today insulin contains 189 bits of information. So how much of this information does insulin require to provide a selective advantage? This question is much more difficult to answer.

This is where human insight is necessary. Some amino acids side chains have very similar chemical properties. Others are similar in size. Thus, some amino acid substitutions should be allowed even if they are not found. These are summarized below with the chemical trait given in parentheses:

Group1: leucine , isoleucine, valine, alanine, and methionine (do not like water, so they tend to cluster on the inside of the protein).
Group 2: tyrosine, phenylalanine, and tryptophan (very large amino acids that can influence protein folding).
Group 3: aspartate and glutamate (acidic side chains, like water).
group 4: histidine, arginine, and lysine (basic side chains, like water).
Group 5: glutamine and asparagine (charged and like water).
Group 6: serine and threonine (like water, tend to be found on the outside of protein).
Group 7: glycine (very small).
Group 8: proline (introduces a bend into the chain).
Group 9: cysteine (cross links peptide chains).

Often these groups can overlap. For example, sometimes a specific site in a protein just needs an amino acid that likes water, and groups 3, 4 and 5 all like water. Sometimes a specific site needs a small amino acid, but it can tolerate amino acids larger than glycine. For simplicity, the model proposed here will not allow the groups to overlap. This makes the hand calculations of knowledge much easier. A more complex model that allows groups to overlap should improve the accuracy, but is not required. This simple analytical model is adequate for most calculations.

Based on these properties, this chapter will propose the following procedure to calculate knowledge:

- If a column in a multiple alignment sequence like figure 4.3 only contains a single amino acid, or if the variation is limited to any one of the above 9 groups, then the column should be included in the calculation for molecular knowledge.

- If the column contains amino acids from different groups, then it should be excluded.

- For columns included in the molecular knowledge calculation, all amino acids in the same group must be included whether they are present in the alignment or not.

For example, at position 19 in table 4.2, only isoleucine and valine are found. But because alanine, methionine, and leucine belong to group 1, it is assumed that these amino acids can be substituted at position 19 without destroying the function of insulin. With this procedure, table 4.2 becomes table 4.4. The parenthesis in table 4.4 represent amino acids that are not present in the multiple sequence alignment (figure 4.3). The positions that are assigned 0 bits have amino acids from more than one of the 9 predefined groups.

Table 4.4: Molecular Knowledge in B chain of Insulin

pos	allowed amino acids	bits	pos	allowed amino acids	bits
2	phe, ala, leu, val	0	17	phe, tyr ,(trp)	3.7
3	val, ala, pro	0	18	leu, (ile),(val), (ala), (met)	1.8
4	pro, lys, asn	0	19	val, ile, (ala), (leu), (met)	1.8
5	gln, (asn)	4	20	cys	5
6	his, arg, (lys)	2.7	21	gly	4
7	leu, (ile), (leu), (val), (met)	1.8	22	asp, glu	4
8	cys	5	23	arg, (lys), (his)	2.7
9	gly	4	24	gly	4
10	ala, pro, ser	0	25	phe, (tyr), (trp)	3.7
11	his, (lys), (arg)	2.4	26	phe, tyr, (trp)	3.7
12	leu, (ile), (val), (ala), (met)	1.8	27	tyr, (phe), (trp)	3.7
13	val, (ile), (leu), (ala), (met)	1.8	28	thr, ser, asn	0
14	glu, asp	4	29	pro	4
15	ala, (leu), (ile),(val), (met)	1.8	30	lys, arg, (his)	2.4
16	leu, (ala), (val),(Ile), (met)	1.8	31	ala, thr, ser, -	0

Total = 76 bits

Example calculation: at position 3 val, ala and pro are found. Because these amino acids are in different groups, the knowledge is defined as zero bits. At position 16 only leu is found, but ala, val, ile, and met probably will not be that damaging to protein function because they are in the same group. The total number of codons that encode these 5 amino acids is 18. Thus, knowledge = 3.32 x log[64/18] = 1.8 bits.

Comparing table 4.4 to table 4.2, it is clear that knowledge is much less than information (76 bits vs. 108 bits). The ratio of knowledge to information for the insulin B chain is thus 76/108 = 70%. The same procedure is repeated for the A chain as shown in table 4.5.

Table 4.5: Molecular Knowledge in Insulin A Chain

pos	allowed amino acids	bits	pos	allowed amino acids	bits
32	gly	4	43	ser, asn, asp	0
33	ile, (val), (leu), (ala), (met)	1.8	44	leu, ile, (val), (ala), (met)	1.8
34	val, (leu), (ile), (ala), (met)	1.8	45	phe, tyr, (trp)	3.7
35	glu, asp	4	46	gln, asp	0
36	gln, (asn)	4	47	leu, (val), (ile), (ala), (met)	1.8
37	cys	5	48	glu, gln	0
38	cys	5	49	asn, ser,his	0
39	glu, his, thr, ala	0	50	tyr, (phe), (trp)	3.7
40	asn, lys, arg, ser, gly	0	51	cys	5
41	pro,thr,ile,val	0	52	asn, (gln)	4
42	cys	5			

Total = 51 bits

Cartoon and Space Fill Models of Insulin

The following pictures of the A and B chains show the location of the amino acids that contain information. The following color code applies: black> 3.4 bits, dark gray> 2.5 bits, gray>2 bits, light gray >1 bit, white <1 bit.

Figure 4.4: Space fill and Cartoon model of B Chain

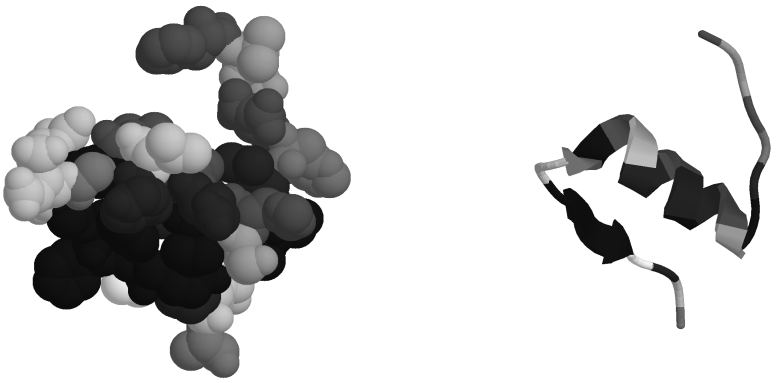


Figure 4.5: Space fill and Cartoon model of A Chain

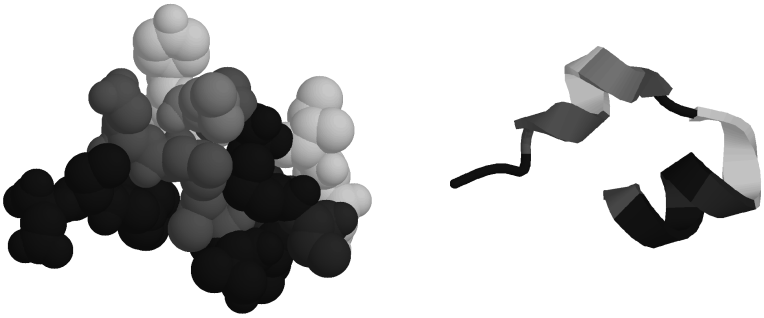
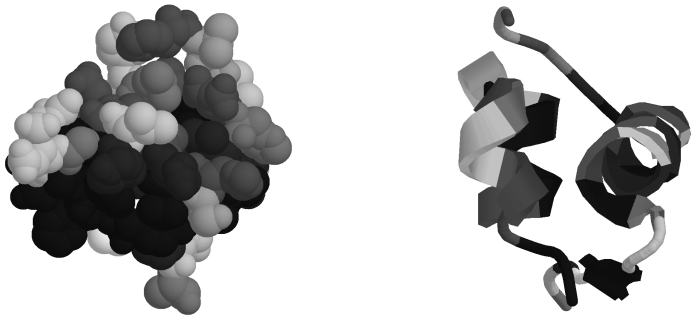


Figure 4.6: Space Fill and Cartoon model of Entire Molecule



The upper right front cover picture also represents insulin. This picture is in color. Red sections contribute > 4 bits per amino acid, orange > 3.4 bits per amino acid, yellow > 2.5 bits per amino acid, and green > 2 bits per amino acid.

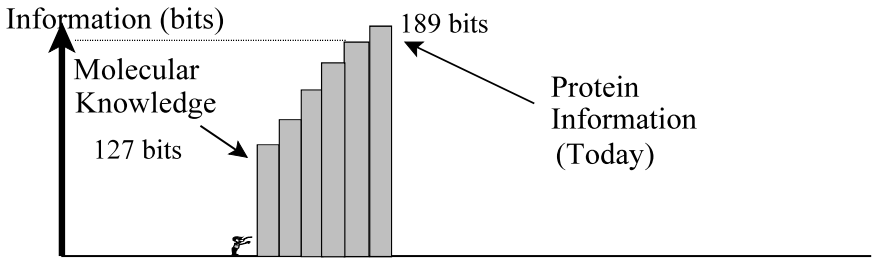
The Probability of Insulin Evolving

The calculation for the amount of information in the insulin A and B chain is based strictly on mathematics. No interpretation is required. The same cannot be said for molecular knowledge. Molecular knowledge requires finding the minimum information that preserves some functionality.

It is a mistake to calculate the odds for the evolution of insulin based on the information in insulin. It is extremely important to use the minimum information that results in a protein with some selective advantage. In other words, molecular knowledge must be used to calculate the probability associated with a protein evolving.

In this chapter, the information in insulin is found to be 189 bits, and the molecular knowledge is calculated at 127 bits. Figure 4.7 illustrates the steps that chance must overcome.

Figure 4.7: Molecular Knowledge in Insulin

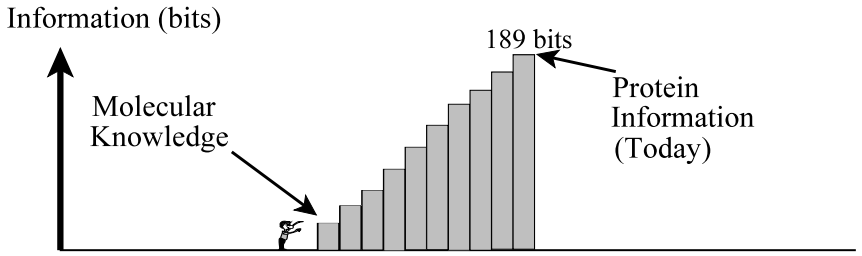


Insulin May Not Imply Design

In most cases, the odds that a protein can evolve are simply 1 in $2^{(\text{molecular knowledge})}$, or in this case, 1 in 2^{127}, but this technique may not apply to insulin. Insulin binds to a protein called the insulin receptor. This receptor senses insulin, and through a few more steps signals cells to absorb sugar from the bloodstream.

This receptor is very specific to the insulin hormone. The original receptor may have been much less specific. So while today insulin requires 127 bits of knowledge, a precursor that might have existed 500 million years ago may have required much less. One could certainly envision a very different insulin receptor. Perhaps this receptor signaled the cell to absorb sugar when it detected any protein greater than 20 amino acids. In this case, the first insulin molecules would have required almost no molecular knowledge, and figure 4.7 might look like figure 4.8.

Figure 4.8: Molecular knowledge in Insulin



If figure 4.8 is accurate then there is a very clear path for Darwinian evolution to work just like Darwin theorized. All of the steps are small; thus, the scientist can easily climb to the top.

Because the structure and specificity of the first insulin receptor is unknown, there is no way to choose between figure 4.7 and 4.8; therefore, the molecular knowledge in insulin cannot be used to reliably infer design. There is no way to choose between figure 4.7 and figure 4.8. Insulin was chosen because it is a very small protein. This makes it easy to manually calculate the information and hence the knowledge. Insulin was not chosen because it implies design. It is merely a convenient learning tool.

Insulin is unique in that its required molecular knowledge depends on its receptor. This allows insulin to co-evolve with its receptor. Most proteins do not have this option.

While figure 4.8 cannot be ruled out, if this figure is an accurate representation, then it can only represent the evolution of insulin before the existence of the common ancestor to fish, amphibians, reptiles, birds and mammals. This must be true because of the similarity of the insulin protein in these diverse species today.

How Accurate is the Technique?

Molecular knowledge should be defined with the following equation:

Molecular knowledge = knowledge per site x number of sites.

Each site is a column in a multiple sequence alignment. So insulin has 52 sites (figure 4.3). The knowledge per site is as follows:

Knowledge per site = $\log_2$ (64 possible condons/ allowed codons).

The big uncertainty in this definition is which codons are allowed. Allowed codons are codons that result in amino acids that yield either a fully functional or marginally functional protein. In other words, an allowed amino acid does not completely destroy the protein's function. If the number of allowed codons can be calculated accurately at each site, then the molecular knowledge so calculated can be related to a probability of evolution. The equation is as follows:

Odds of a protein evolving = 1 in $2^{\text{(molecular knowledge)}}$ chance.

The error in this technique arises because it is not easy to determine which amino acids are allowed. The number of allowed amino acids is not the number of amino acids found in a particular column in a multiple sequence alignment. Natural selection tends to weed out functional proteins that are slightly deleterious. This means that in many columns the number of amino acids that are allowed is significantly greater than what is observed in nature.

There are two ways to determine the allowed amino acids. One is through experimentation. The other is to develop a set of rules that allow additional amino acids at each site in a protein. This chapter proposed a set of rules that grouped amino acids into 9 groups and proposed a method to determine the true number of allowed amino acids at each site. Since these rules are arbitrary, the most pressing question is do these rules accurately predict allowed amino acids?

At first glance, the rules seem too stringent. For example, at position 24 in table 4.5 only valine is found. With these rules, leucine, isoleucine, methionine and alanine are also allowed because they belong to the same amino acid group. Experimental evidence shows that if this valine is replaced with leucine, the resulting insulin will lose more than 90% of its functionality.[2] Replacing it with an amino acid from a different group should further destroy functionality (leaving the insulin molecule with almost no functionality). Positions 25 and 26 of the B chain are always tyrosine or phenylalanine. Replacing position 26 with a serine almost completely destroys insulin functionality.[1] Position 25 is not as critical because a serine here only destroys 85% of insulin's functionality.[1] A single amino acid substitution can in many cases result in a non-functional protein.

While no analytical technique designed to calculate molecular knowledge will be perfect, the one introduced in this chapter in most cases should be close to the true value of molecular knowledge. The only way to validate this analytical technique is to compare its predictions to experimental evidence. Researchers can directly substitute amino acids at each site in a protein and then screen the resulting protein for functionality. From this experimental evidence, it is possible to directly calculate molecular knowledge.

Appendix 6 compares the analytical techniques developed in this chapter to experimental results, and these comparisons support figure 4.9. In this figure, the actual knowledge is calculated from the experimental techniques described in appendix 6. The knowledge predicted by the analytical technique agrees very well with the experimental data, but the two calculations are seldom identical. Sometimes the analytical technique yields more knowledge and sometimes it yields less. The shaded region in figure 4.9 represents this uncertainty. The agreement between the two approaches could be improved with a more complex set of rules to better estimate the allowed amino acids. Nevertheless, even with the simple set of rules proposed in this chapter, the calculations agree remarkably well.

Figure 4.9: Actual knowledge vs. Assigned Knowledge

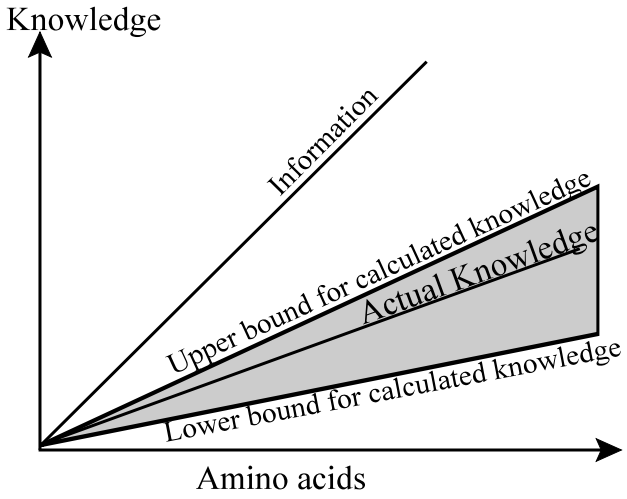
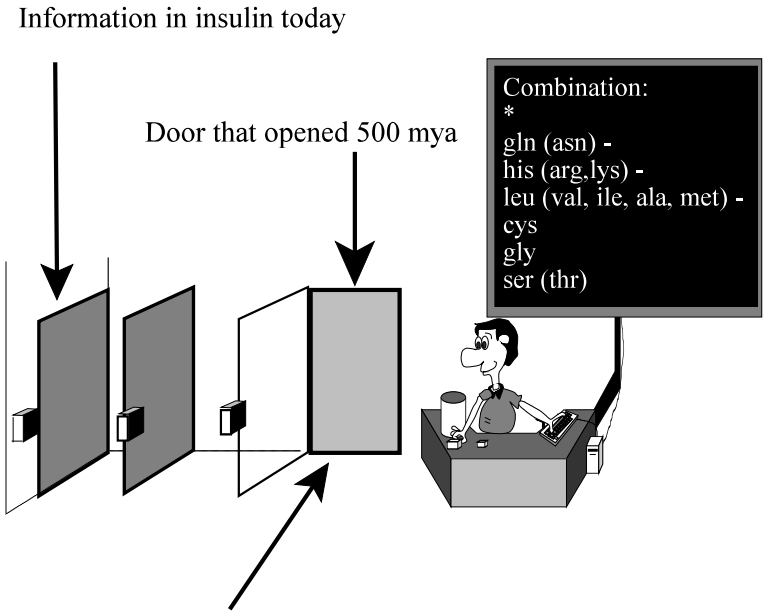


Figure 4.10 on the next page should be compared to figure 4.1 at the beginning of this chapter. In figure 4.1, the technique used to calculate the information of insulin (as it exists today) is shown. This information cannot be related to a probability because insulin has already been optimized by natural selection. In figure 4.1, the only door is the last door. All of the doors leading up to this door are hidden. The technique introduced in this chapter attempts to reconstruct the combination of the earlier doors. In particular, the first door is important because it is this door that determines whether or not naturalistic laws can explain the evolution of insulin. The screen in figure 4.10 shows the combinations that will open the first door. For example, the first position is represented by an asterisk because all 20 amino acids are allowed at this position. In the second position, only gln is found today, but since asn belongs to the same group, asn is shown in parenthesis. Any combination with either gln or asn at the second position will open the door. The accuracy of this technique depends on how well it predicts the combination of the first door. Insulin may not imply design for reasons already discussed. Nevertheless, when other proteins are analyzed with this method, the design inference is very strong.

Figure 4.10: Molecular Knowledge of Insulin



Experimental Evidence for figure 4.9: In appendix 6, the techniques developed in this chapter are compared to experiments in which researchers randomly mutate proteins and then screen for functionality. Reading appendix 6, immediately after finishing this chapter is encouraged because the two subjects complement each other well.

.

Procedures Needed to Reproduce this Data:

The amino acid sequences of insulin and almost all other proteins found in life are available at http://us.expasy.org/sprot/. A search for insulin on this web site returns the sequence for insulin in many different species. Browse down the search results to locate INS_HUMAN. This link contains the sequence for human insulin at the very bottom of the page.

Figure 4.3 was constructed by downloading the insulin amino acid sequence for 18 species in FASTA format, and then running the alignment program, ClustalX, which is available at this URL: http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html

The sample size was limited to 18 species because the online databases are not without error. The criteria to include a position in molecular knowledge is very sensitive to these errors. For example, suppose that position 10 in a specific protein is always a glycine, but due to an error in the database, one of the entries reports that a valine is found at this position instead of glycine. This error means that position 10 must be excluded from the calculation in molecular knowledge. So to make this technique less sensitive to database errors, the sample size is arbitrarily limited to 18 entries. The diversity of the database is very important. The more diverse the species, the better this technique works.

References:

1) Shoelson et al.,"Identification of a Mutant Human Insulin Predicted to Contain a Serine for Phenylalanine Substitution," Procedding of the National Association for Science, Dec 1983, vol 80, 7390 -7390.

2) Sakura H., et al, "Structurally Abnormal Insulin in a diabetic Patient Characteristic of the Mutant Insulin A3 (Val-> leu) Isolates from the Pancreas," J. Clin. Invest. 78:1666-1672, 1986.

3) The Neutral Theory of Evolution, Kimura, 1983.

4) Thompson, Gibson, Plewniak, Jeanmougin and Higgins, The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882, 1997.